CREATION OF DISTRIBUTED SERVICE SYSTEM DEMAND SURFACES TO INFORM DESIGN DECISIONS IN NOVEL SCENARIOS

A Thesis
Presented to
The Academic Faculty

By

Bryan Calvert Watson

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in Mechanical Engineering

Georgia Institute of Technology

August, 2019

CREATION OF USER DEMAND SURFACES TO INFORM DESIGN DECISIONS IN NOVEL SCENARIOS

Approved By:

Dr. Cassandra Telenko
Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Julie Linsey
Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Bert Bras
Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: 25 APR 2019

This work is dedicated to all those who have supported me, most of all to Ashley and my family.

ACKNOWLEDGEMENTS

Special thanks to my Thesis committee for their support and insight, the fellow members of the CASS Lab, and DIVVY for publishing their data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS, LIST OF ABBREVIATIONS

Variable	Symbol
Mean Ridership	μ
User Population Size	*n*
User Population Product Affinity (Probability of Use).	*p*
Average Number of Riders	*AVGn*
Average Ride Probability	*AVGp*
Bike Path Length	*BikePath*
Shared Bike Path	*ShBP*
Separate Bike Path	*SeBP*
Number of Restaurants	*Food*
Number of Retail Locations	*Retail*
Number of Bike and Pedestrian Crashes	*BikeCrash*
Number of Park Attractions	*Park*
Number of Train Stops	*Train*
Distance from Center of Business District	*DISTcbd*
Distance from Lake Michigan	*DISTlm (DISTH2O)*
Number of Stations within 4800 ft	*St<4800*
Distance to Nearest Station	*NearSt*
Total Population	*Pop*
Median Income	*$*
Percent Caucasian	*%C*
Population 6-64	*6-64*

Variable	Symbol
Number of Alternate Commuters	*AltCom*
ATM	*ATM*
Parking Garages	*PG*
Bus Stops	*BUS*
Elevation	*Alt*
Group Quarters	*Group*
Bike Friendly Road	*BikeRoad*
Employment	*Employ*
Campus	*Campus*
Informal Path	*IP*
Tourist Attraction	*Tours*
Services	*S*
Land Use Mixture	*LandMix*
Percent Public Transit	*%PT*
High Income	*$$*
Crime	*Crime*
Hotel	*Hotel*
% Drivers	*%D*
Type of Road	*TRoad*
Number of Low Vehicle Individuals	*lowV*
Number of Bachelor's Degrees	*Bach*

SUMMARY

Traditional demand estimation tools were developed for product design instead of product service system (PSS) design. PSS is a new market structure where the focus is on selling the use of a product instead of the product itself. Demand estimation faces challenges when applied to PSS design including mis-estimation, not being quantifiably repeatable, or built from evidence.

This thesis examines two PSS Design methodology questions. First, *what is the effectiveness of spatially-derived revealed preference data in estimating distributed PSS demand?* Estimating binomial distribution parameters *n* (user population size) and *p* (user population product affinity) can predict demand in new situations for distributed product service systems. Plots of binomial parameters reveal a continuous surface over the PSS area that allow more accurate prediction of relative ridership levels at new PSS locations.

Secondly, this work examine *how designers can compensate for situations where the PSS design environment has changed and limited user data is available to create demand estimations*. This thesis hypothesis that publicly available socio-demographic and environmental variables can inform multivariable regressions that estimate the *n* and *p* Demand Surfaces outside of the boundaries previously constrained by available user data.

Together, the answers to these two questions provide an initial framework to estimate Revealed Preference demand for many types of PSSs. In the examination of both questions, the proposed approaches are tested by the 2015 Chicago Bike Share System expansion. The effectiveness of these approaches is shown through analysis techniques including Spearman's rho, Pearson's Coefficient, Monte-Carlo Sensitivity Analysis, and resource impact of algorithm implementation.

CHAPTER 1: Estimating User Demand to Design Product Service Systems

1.1 Demand Challenge Facing Product Service System Expansion

Product service systems (PSSs) face a unique problem slowing their adoption and growth: their performance depends upon accurate demand estimation approaches, yet current demand estimation methods were developed for individual products. PSS is a new market structure where the focus is on selling the use of a product instead of the product itself [1,2]. Many of these PSSs rely on managing distributed services, such as providing vehicles at point of service or distributing and recollecting clothing or tools. Bike-share, car-share, and renting complex portions of aircraft, such as Rolls-Royce's selling turbine 'power-by-the-hour' to airlines are all examples of PSSs [2]. PSSs provide benefits such as improving environmental sustainability without reducing product features [1,2]. For example, one model estimates that PSS implementation of washing machines could lower $C0_2$ emissions by approximately 10%, reduce fluctuations in supply chain demand, and reduce the number of machines in service[3].

User acceptance of the PSS is key for success. Recognizing this importance, previous efforts incorporate user data and preferences into the design process with the goal of improving the final user experience [4]. *User-oriented design, User-centered design,* and *Usage context-based design* incorporate user desires and experiences to inform the creation and functionality of new products [4–6]. User data drives the design process rather than choosing technology and functionality based on precedents or assumptions about user needs or responses [4]. Methods such as user interviews, role playing, or user interactions with prototypes are often used to extract user data [5]. Additionally, customer experience with similar models can allow a basis for updated designs [7]. These techniques for user-oriented design can result in products with higher adoption

rates[4]. Conversely, misestimating user demand can result in financial consequences for the firm in error [7,8].

Approaches such as Stated Preference (SP), Revealed Preference (RP), and Stated Intention (SI) utilize a combination of quantitative and qualitative methods to estimate user demand [9]. Additionally, it is unlikely that user demand for a PSS is equivalent with user demand to purchase a product. This difference could lead to high error rates because most PSS demand estimation methods are an extension of product demand estimation methods [1,2]. Local knowledge, expert guidance, heuristics, or "gut-feel of the decision maker" may also be employed to estimate demand [7,10,11]. As a result, design decisions may lead to mis-estimation, and they may not be quantifiably repeatable or built from evidence [3]. This may indicate the need for new methods of PSS demand estimation [1]. Additionally, many approaches focus on modeling currently existing demand patterns, rather than predicting demand patterns in new situations, providing limited use for PSS expansion planning.

1.2 Current Approaches to Product Demand Estimation and Modeling

Previous research has used a variety of approaches have been to attempt to overcome these challenges and improve PSS design. Reviews of 80 emerging market case studies were used to generate 9 decision making heuristics for PSS design in developing countries [9]. Beyond general heuristics, another challenge is translating imprecise customer responses into PSS design characteristics. Different individuals may describe the desired PSS differently due to different priorities or speech patterns. One approach to this problem, Song et. al. introduced Industrial Customer Activity Cycle (I-CAC) Analysis, providing an improved method to translate customer requirements into a PSS design [10]. A second approach used Supervised Machine Learning to estimate PSS design configuration from data, achieving an impressive classification accuracy of

93.3% [11]. These studies however, look only at overall PSS design, not utilization of local demand variance within the PSS to inform design. Once a general PSS framework has been chosen, an approach is needed to inform the implementation of the PSS which is sensitive to local demand fluctuations.

There are three major approaches to product demand estimation: Stated Preference (SP), Stated Intention (SI), and Revealed Preference (RP). SP surveys ask the potential user to state which configuration or design they would prefer, allowing for demand estimation of new products and services. SI investigators ask users to state their intended use of the new product or system. User data is collected via surveys, crowdsourcing designer knowledge, and focus groups [9,10]. These methods could be susceptible to small study effects, selection bias, response bias, gamification, or confusing survey wording [7,9]. For example, the number and quality of crowdsourcing solutions may be dependent on the number and size of prizes offered for the contest [12]. SP and SI methods are also susceptible to the user's inability to accurately forecast their demand for a new service and have been shown to systematically overpredict demand [9]. SI is susceptible to self-selectivity bias and gamification by users who desire service expansion [9]. SP and SI surveys require significant training, time, and other resources to implement well and may be poorly implemented in practice [7]. The collection of SI and SP surveys can also create a large financial or logistic cost to an organization. A study to predict train demand distributed 29,873 SI surveys and 1,254 SP surveys [9].

As an alternative or complement to SP and SI, RP demand estimations may be created by observing user demand for comparable products [13]. Observations provide information about desired design features, but some respondents may make choices habitually without considering all of the alternatives [9]. As a result, some studies include "noncompensatory behavior", such as

heuristics, to initially limit the options considered into the user models [14]. Additionally, for design variables that are codependent, RP may not identify which design variable stands in causal relation to the observed user demand. For example, in rail transit travel time and cost are positively correlated [9]. For a completely novel product, RP approaches are more difficult; designers may consider creating a test market to observe users, but this may not be financially viable [7].

Despite these limitations, SP, SI, and RP have been successfully implemented in multiple product demand estimation models. Frischknect et al. analyzed RP data in an economic model approach where potential users desired to maximize their utility, which was modeled as a function of product features. The resulting model, however, indicated that the introduction of an identical product would increase the total market size [15]. He et al. created a framework for Usage Context-Based Design where the user demand is a function of the user characteristics, usage context, and product characteristics. During data collection for case study models, SP and SI techniques are utilized. He et al. demonstrated their approach for both a jigsaw and hybrid electric vehicle [6]. Contrasting with the two prior approaches, William et al.'s investigation focused on the influence that large retail stores have on user demand. Their model incorporates retail preference and limitations to this profit maximization problem due to limited retailer shelf space [16].

Demand models may be replaced with expert-generated heuristics [7] or otherwise defined by an expert. One study used agent-based modeling to inform design decisions for a cordless angle grinder. The demand model focused on the relationship between design characteristics and competing manufacturers choices to understand the overall market dynamic [17]. Rather than estimate user demand, this study assumes that "customer preferences are assumed to be common knowledge to the firms [17]."

Algorithms have also been developed for demand estimation, but these do not focus on Distributed or PSS demand estimation. These algorithms fall into several categories. First, many focus on identification of desired physical product attributes [18,19] and are not tested or developed for distributed systems. Chen, Honda, and Yang utilized machine learning approaches to supplement Stated Preference identification of significant attributes of solar panels [18]. A second area investigates and develops discrete choice models to predict market level demand for products [20,21]. This thrust is also not developed for distributed systems. For example, Haaf et.al examined the sensitivity of market share predictions to changes in the applied Discrete Choice Model [20]. Finally, a third focus is on distributed systems, but investigates distributed system network connections and dynamics [22]. This third thrust does not focus on differences in user population demand within the system, rather it focuses on understanding and predicting the way in which these systems form and organize. For example, Sha and Panchi provide methods to examine the decision making behavior of local nodes in an internet case study. Unlike PSS, the decisions of these individual nodes to connect and disconnect define the structure of the system.

PSS demand estimation faces additional unique challenges. First, using traditional tools may result in inaccurate results especially when evaluating service expansion. In one notable service expansion study, the results of 29 firms' wire-less demand estimations were reviewed. The technology was new, so firms could not use previous user demand as a guide and creating a mockup or test market was financially infeasible. Multiple approaches were utilized: two estimates based on radio-paging demand, one model based on demographic factors, and 12 on expert judgement or heuristics. Additionally, 27 firms utilized market surveys with 7 of the 19 questions used being SI. The authors noted the following concerns with the methods used: the surveys had issues such as confusing wording and leading the respondent, simple statistical methods were used, and the firms generally ignored non-response surveys in their approach. As a result of the divergent approaches taken by the firms, the market estimates varied from 14,000 to 91,000 expected subscribers for the same service area. The authors conclude that the modeling approaches require additional refinement, potentially by combining multiple demand estimation sources such as expert guidance, surveys, and demand models [7].

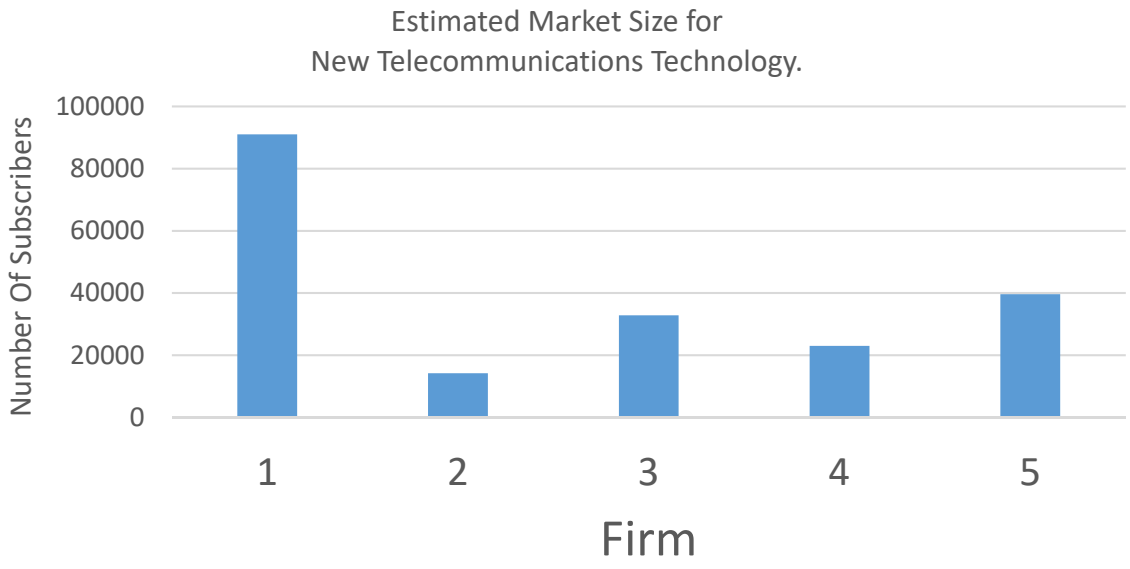


Figure 1: Example Variance in PSS Market Size Estimation[7].

The second challenge faced by PSS demand estimation is the need to validate the accuracy of currently applied demand estimation tools [1,2]. For example, one study utilized 245 SP surveys to estimate user demand for a model informing autonomous sharing vehicle infrastructure design, but no assessment was performed of demand prediction accuracy [8,23]. PSS demand estimation difficulty may be due to the sensitivity to both system level influences, changes in individual agents' demands, or the influence of demand timing on product flows [2,24]. Further complicating these approaches is the uncertain value of user time, necessary for RP estimation, up to 100% variation [9]. Finally, SP or SI PSS data requires extensive surveys which may be financially limiting for the designer or operator. Due to the geographically distributed nature of PSS, more individuals may need to be surveyed than for product demand estimation.

1.3 Proposed PSS Demand Estimation Framework

Thus, PSS expansion and adoption faces a crucial problem. Accurate Demand estimation approaches are required to maximize performance and adoption of PSS, but current SP, RP, SI demand estimation approaches require improvement when applied to PSS design. This thesis proposes a framework for resolving difficulties creating RP-based demand estimates in new situations in distributed PSS. The approach utilizes available user data to create a probability map that describe user characteristics related to the current demand and provides a means to predict future demand at new locations within that space. This framework would provide designers additional needed information when making design decisions in situations such as product expansions.

Specially, this framework is applicable within two scenarios. First, how can designers make the best decisions when user demand information is available? I propose that estimating binomial distribution parameters n (user population size) and p (user population product affinity) can predict

demand in new situations for distributed product service systems. Plots of binomial parameters reveal continuous demand surfaces over the PSS area that allow more accurate prediction of relative ridership levels at new PSS locations when undergoing an increase in system density. This application is tested in Case Study One.

In the second scenario, I examine how designers can compensate for situations where the PSS design environment has changed and limited user data is available to create demand estimations. The results show that publicly available socio-demographic and environmental variables can be used with multivariable regressions to estimate the n and p Demand Surfaces outside of the boundaries previously constrained by available user data. This is applicable for a PSS that increases beyond the current system boundaries.

This investigation of estimating user demand in new situations for distributed PSS falls within two literature gaps. First, within the demand field, demand estimation algorithm research is focused on individual product demand [18–21], not distributed system demand. Secondly, distributed system research often focuses on algorithms that address operational concerns for redistributing products within a system [25–28], while this research provides a tool for the initial design phase. The approach in this thesis utilizes available user data to create a surface that describes user characteristics n and p. This surface is used to predict future demand at new locations within that space and provides designers needed information when making design decisions in situations such as product expansions.

The contributions of this thesis can be grouped by the two central questions examined in this thesis. In exploring the first question (*what is the effectiveness of spatially-derived revealed preference data in estimating distributed PSS demand?)* the following contributions are made:

1. The development and application of a revealed preference demand estimation method for distributed product service systems.
2. The creation and application of novel geo-spatial demand surfaces from user data.
3. A tool is created and tested to improve demand prediction for PSS undergoing an increase in system density.

Secondly, when examining *how designers can compensate for situations where the PSS design environment has changed and limited user data is available to create demand estimations*, the following contributions are made:

1. We provide a framework to transform geographically limited available PSS user data into design insights for the portion of the system without user data.
2. A second validation is conducted of the value of *n* and *p* estimations for PSS planning as proposed in the first question examined.
3. We provide a tool for PSS operators planning a system expansion.
4. We identify environmental and socio-demographic variables that correlate with higher Bike Share System usage.

1.4 Socio Technical Environmental Systems: Applied Modeling Perspective

This work approaches the development of this framework from the perspective of Socio-Technical-Environmental Systems thinking (STES). STES are systems defined by the complex interaction of technical artifacts, intelligent agents, and the environment they exist within. Technical artifacts are those objects created by humans designed to achieve specific goals [29]. Intelligent agents are the humans who interact with the technical artifacts whose decisions must be considered. Both systems interact in a changing environment. For example, consider the traffic around a busy traffic circle. The flow of cars through this system is defined by the type of vehicles

(technical artifacts), choices of the drivers (intelligent agents), and the road conditions due to weather (environment). Although some have proposed expanding Socio-Technical Systems (STS) to incorporate some of the elements described here[30], I find that a distinct approach has the advantage of avoiding linguistic confusion or importing preconceived notions regarding STS. This thesis is an expansion of the efforts made in socio-technical systems (STS), distinct in purpose and boundaries.

STS thinking originated in the 1950s as a means to increase industrial productivity by understanding and improving the social and technical aspects of the system [31]. Subsequent applications have reflected this initial goal, including forays into coal mining, textile mills, merchant shipping, occupational safety, or defense parts acquisition [31–33]. STS often utilizes social organization or theory as a solution to organizational issues. In contrast to this goal of improving economic productivity, STES thinking builds upon the STS principle that the technical artifact is critical to determining the performance characteristics of the STES [34]. As design engineers, we seek to use insights about the interaction between the technical artifact and the STES to make design decisions. STES could meet a variety of goals beyond economic productivity. These goals range from improved system flow, environmental sustainability, and others. For the traffic circle example, goals might include changing car speed to maximize traffic flow or minimize environmental impacts.

Perhaps the clearest distinction between STES and STS is the lack of an environmental boundary within STS. This is somewhat misleading. Foundationally, STS is applied to systems where the environment is approximated as steady state [31]. This is because STS founders believed that the organization was able to control the environment through technical or social interventions. For many STS applications, environmental factors outside of human or

technological systems do not appear to be applicable. Consider the rearrangement of a workforce in a textile factory [31]. This system is entirely enclosed within the factory walls, any change in temperature or humidity is due to management choices about climate control. Thus, STES define the environmental system as environmental systems beyond the intelligent agent's ability to manipulate. Considering the example of the traffic circle, the temperature inside the cars would not be considered part of the environmental system, while the rainfall or fog outside the car would be considered environmental.

Secondly the scale of STES is distinct, particularly due to the aggregation of intelligent agents within STS. Like STS, STES assumes agents act with varying degrees of autonomy and that they are able to learn[31]. Many STS studies however, incorporate the individuals into groups as a key method of analysis[31]. Although group aggregation can be used within STES, the problems of most interest is the MESO (human) scale. STESs exhibit emergent system-level characteristics due to intelligent agent self-organization [32,33,35]. Emergent characteristics are system-level patterns that are not predictable from reductionist approaches [32].

By considering demand estimation of PSSs from this framework, designers can seek to model all three necessary influences on systems performance: human agents, environment, and technical artifacts.

1.5 Potential Appropriate Mathematical Models

Successful modeling of a STES requires the identification of the underlying mathematical model. Product or service utilization can be understood as aggregated individual choices. This framework supposes the following:

1) The events are independent from one another.
2) The total events which are observed occur over a finite time frame.

3) Each event consists of an individual making a binary choice.

This framework is also applicable to other collections of binary events, such as machine failures, instances of disability, or success rate of capturing wild animals [36–38].

The appropriate mathematical distribution to describe this system is a discrete function. Individual choices are either success or failures and the total observed successes in each dataset will be an integer. Candidates are the binomial, negative binomial, and Poisson distribution.

The primary difference between the Poisson and binomial distributions lies in the number of trials observed. The binomial distribution approaches the Poisson distribution as the number of observed trials increases and the probability of success for each trial decreases [39]. Additionally, the Poisson distribution assumes that the population mean and variance are equal, which may not be true for the analyzed system [38].

The key difference between the applicability of the negative binomial and binomial distribution is the information necessary to define each. The binomial distribution is defined by the number of trials, n and the probability of success for each trial, p. In contrast, the negative binomial distribution is defined by the number of successes x and the probability of success p. The binomial distribution can be used to yield the number of successes, x, while the negative binomial can be used to determine the number of trials required to observe a certain number of successes.

When fitting the negative binomial distribution to data, authors can observe x and solve for p. The binomial distribution, however, may require estimation of both n and p. Although more complex, the estimation of n and p provide additional information concerning the system being examined.

1.6 Interpretation of Individual Choices as a Binomial Distribution

Once the appropriate mathematical model has been identified for the framework, it should be scrutinized for applicability to this design problem. Consider the case of a single individual making a decision to utilize a product. For this discussion the term "product" refers to any resource consumed by an individual including transportation infrastructure, consumables, or the service industry. This binary event can be evaluated as a Bernoulli Distribution.

$$f(Ri = x|p) = p^x(1-p)^{1-x} \tag{1}$$

The individual (*Ri)* could select either to utilize the product (x=1) or not (x=0). This distribution is defined by a single parameter, p the probability that the individual will choose to utilize the product.

Now, expand this case to the scenario of a queue of total length n, observed over some finite time period. This scenario is simply a repeated Bernoulli distribution, defined as a binomial Distribution. This distribution yields the number of successes (x), defined by two distribution parameters: n and p.

$$f(x|p) = \binom{n}{x} p^x(1-p)^{n-x} \tag{2}$$

As previously stated, n is the number of trials which occurs over the finite time period. p remains the probability of success for each trial. This interpretation of n and p is consistent with the literature's *a priori* assumption that n and p are independent [37]. Additionally, n and p are considered to be fixed, but unknown [39]. Observations over multiple comparable finite periods

of time creates a dataset, such that X_i (i= 1, 2, … r) [40]. Where X is independently and identically distributed from Bin(n,p).

The significance of understanding total consumer usage as multiple binomial distributions is that n and p encode additional information unavailable from other methods. For example, consider the case where a designer chooses to use mean utilization of a current product to estimate new product usage. As shown in Eq. (3), multiple n, p combinations exist that result in the same expected value.

$$E[x] = n * p \tag{3}$$

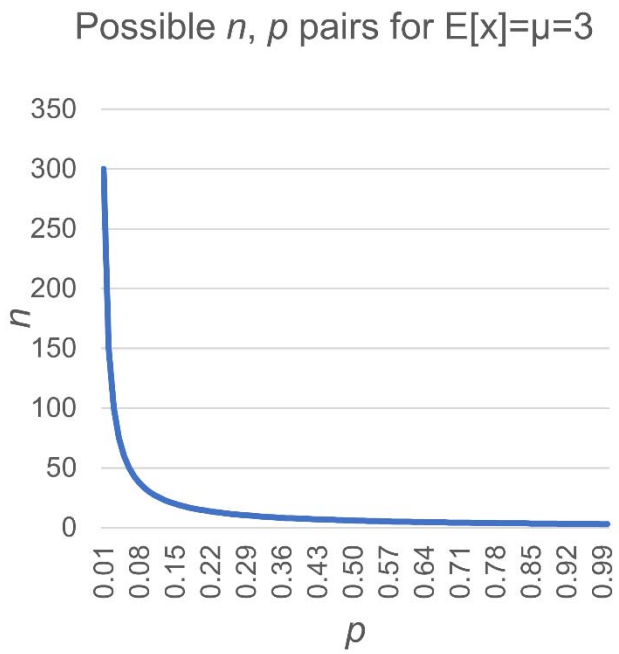


Figure 2: Possible n, p combinations that yield E[x] =μ=3

Figure 2 illustrates that a high number of possible consumers (n) who are each unlikely to utilize the service (p) would yield the same E[x] as a condition with high population affinity (p) and low population (n). Thus, determining n and p could provide additional, critical information for design decisions. For example, a decision maker might observe that a certain PSS location has an average of 3 rentals per hour (E[x]=3). This could be because the location has a high n

(population of potential users), but each user has a low probability of riding (*p)*. The opposite could also be true. Knowing which design situation the PSS decision maker is in informs intervention strategies. For example, for a situation with a low *p* the decision maker might choose to make riding more attractive by lowering the price to ride. For a situation with a low *n,* however, one could not expect lowering the price to change the number of potential users, only reduce the profit generated by that location.

1.7 Case Study Selection for Framework Validation

Now that the appropriate mathematical approach has been identified, a compelling case study can be used to test the effectiveness of our approach. One area where this type of tool could be immediately applied is within the rapid growth of Bike Sharing Systems (BSS). BSS are currently utilized in over 800 cities on five continents [41]. BSS often rely on bikes being placed around the city in docks at prescribed locations, awaiting patron checkout or return [42]. As BSS demand increases, BSS operators must make key design decisions about the location and size of new BSS docks. Demand predictions also affect BSS pricing schemes which usually charge a very low or no cost for the first portion of a ride, followed by accumulating rental fees [41]. Current guidance exists for new BSS dock location such as methods to determine BSS dock density or typically attractive locations, but guidance for BSS dock size is limited to SI and SP methods [10,43].

Design decisions determining BSS dock size play a critical role in both the user experience and sustainability of the BSS scheme. These design decisions influence the physical characteristics of the BSS docks (dock size) and thus directly influences customer satisfaction. Multiple studies have shown that convenience is a key reason people utilize BSS [27,44,45]. If BSS docks are too small, users may become frustrated when they are unable to locate a bike to check-out or an empty dock for bike return. Users who become frustrated because of lack of availability (underestimated

demand) may cease using BSS, also diminishing operator revenue. If user decline is significant enough, BSS may no longer be economically sustainable for the operators.

Conversely, BSS docks that are too large for the existing demand can unnecessarily increase monetary and environmental costs of the service. BSS docks can each require approximately 30.4 kg (67 lbs.) of steel [46]. Unnecessarily increasing the BSS dock sizing to ensure all demand is met will threaten both economic sustainability due increased BSS operator costs and also the underlying environmental sustainability. BSS docks also often replace existing parking or traffic lanes. Thus, large unused BSS docks could be met with public resistance.

Operators could resize existing modular BSS stations to match demand [10], but face several challenges. The first challenge is that obtaining measurements of rider frustration or inability to dock require additional time and resource investment beyond the initial expansion. Additionally, dock power supply requirements may render the approach of modifying existing stations ineffective [43]. These unmet design needs prompt a key question: *How can new BSS Dock usage be estimated prior to implementation using a repeatable, quantitative method, minimizing new station sizing inaccuracy?* Or, more generally: *What is the effectiveness of spatially-derived revealed preference data in estimating distributed PSS demand?*

This work hypothesizes that RP-based demand estimation can be achieved using a binomial distribution. Parameters of the binomial estimated from historical data can be utilized to predict potential ridership for new stations. These parameters can be uniquely determined for each existing station and used to predict ridership levels at new station locations. These generated Demand Surfaces could effectively replace or supplement current approaches such as BSS staffing knowledge or origin-destination surveys. Plots of these parameters reveal a continuous surface

over the area serviced by the BSS. This approach is validated by utilizing the data from a major BSS expansion.

The data is evaluated in the form of two case studies. Case Study One evaluates the scenario where a PSS expands by increasing density. In this scenario, user data in the same environment is available to estimate the *n* and *p* surfaces. This application, however, requires user data to be available in the PSS expansion environment. Case Study Two explores that gap, developing an approach to estimate *n* and *p* in environments without user data. This scenario would be applicable where a PSS expands beyond its current boundary. Together, these results provide a framework to estimate RP demand for many types of service or product expansions. The goal is that this framework be used by designers to provide additional information to improve their decision-making process during a product or service expansion.

1.8 Summary

Chapter 1 described the unique PSS design challenge: the need for more accurate PSS demand estimation methods. Current demand estimation approaches were described. The concept of Socio-Technical Environmental Systems and its relation to Socio-Technical System modeling was discussed to explain our modeling philosophy. Finally, the binomial distribution was presented, followed by interpretation of parameters *n* and *p*. Arguments were presented that determining *n* and *p* provide additional information about system demand than hourly demand (E[x]) alone.

The remainder of this thesis is organized as follows. Chapter Two presents historic attempts to estimate *n* and *p* followed by the method chosen in this study to estimate *n* and *p* from available RP data. Monte Carlo Simulations are then presented to validate the *n* and *p* estimation approach.

Chapters 3 and 4 examines the first question this RP framework examines: *What is the effectiveness of spatially-derived revealed preference data in estimating distributed PSS demand?* The approach of estimating *n* and *p* surfaces from historic RP data is applied to a Case Study of the 2015 Chicago BSS expansion. Chapter 3 presents necessary background information to fully understand the methodology employed in Case Study One. First, to increase understanding of the role improved RP demand could play in the BSS industry, current station sizing methods are presented and described. Then, current BSS modeling work is presented to both highlight the difference between this study and previous efforts and provide evidence for methodological choices. Chapter 4 presents and discusses the two results from Case Study One. First, the actual generated *n* and *p* surfaces themselves were analyzed for insights. Secondly, the accuracy of the algorithmic predictions is compared to the implemented operator ordering. The results show that Case Study One demonstrates a new RP demand estimation method to assist designers planning product or service expansion, providing critical design data for PSS infrastructure expansion.

Chapters 5 and 6 examine the second question proposed by this thesis: *how can designers compensate for situations where the PSS design environment has changed and limited user data is available to create demand estimations*? This examination seeks to generalize the approach shown in Case Study One to situations where RP data is not available. Chapter 5 contains background and methodology for Case Study Two. Background provides historical insight into the problem estimating demand in situations without user data and previous BSS regression attempts. Methodology presents the data sources utilized for the regression, a brief description of the regression approach taken, and the four tests used to validate this approach. Chapter 6 presents the results of Case Study Two including the generated regressions and test results. The results

show that this approach provides a better demand estimate for stations outside the boundaries of available RP data than traditional demand estimation methods.

Finally, Chapter 7 provides a summary of the significant findings of this work and discusses opportunities for future work. This includes opportunities for future investigations and consolidated design recommendations for PSS decision makers.

CHAPTER 2: Simultaneous Estimation of Binomial Parameters *n* and *p*

Chapter 1 introduced the idea that determination of n and p could provide additional information about Product Service System (PSS) Demand. This chapter presents historic attempts to estimate n and p followed by the method chosen in this study to estimate n and p from historic user data. Finally, Monte Carlo Simulations are presented to validate the chosen n and p estimation approach.

2.1 Current Methods to Determine *n* and *p*

Simultaneous determination of both n and p has been considered a classic statistics problem [47]. These investigations have been motivated by the desire to derive underlying population characteristics from observational data. Olkin, Petkau, and Zidek discuss the analysis of crime data with unreported offenses [39]. Other scenarios include estimating the total number of appliances in an area given an observed repair rate [37]. Efforts thus far recognize that approaches generally underestimate n and that estimates of n tend to wildly fluctuate with small sample changes [47]. For a more complete review the reader is referred to an introduction written by DasGupta and Rubin for a summary of previous efforts [47].

The most straightforward approach to estimate n and p is the method of moments which utilized the relationships shown in Eq. (4) and Eq. (5) to calculate n and p. S_x is the sample set X_i (i= 1, 2, … r) variance. E[x] is the expected value of Bin(n,p), while μ refers the mean of the observed sample set X_i (i= 1, 2, … r).

$$E[x] = \mu = n * p \tag{4}$$

$$n * p(1 - p) = S_x \tag{5}$$

Eq. (6) shows the relationship between the expected utilization during a finite period of time (*E[x]),* the utilization observed (μ), *n,* and *p. r* is the number of datapoints.

$$E[x] = \mu = \lim_{r\to\infty} \frac{1}{r} \sum_{i}^{r} X_i \tag{6}$$

The method of moments has also been updated with new identities using observed sample maximums that can be shown to asymptotically approach *n* [47]. This approach shows promise for small *p*, unless *n* is also small. Maximum Likelihood approaches have also been investigated, but they also produce unstable estimates of *n* and *p* [47]. The erratic behavior of Maximum Likelihood and Method of Moment estimators is because they can exhibit heavy-tailed, non-normal distributions [48]. This erratic behavior can cause a small sample change to result in a large change in the estimated value of *n*. For MLE this erratic behavior is due to the fact that when σ^2/μ^2 is near one, the likelihood function is essentially level [39]. σ^2 is the population variance. Hall also shows that this erratic behavior is inversely proportional to the magnitude of rp^2 [48], highlighting the importance of large sample sets (*r)*. Other solutions include recognizing that the maximum number of successes observed asymptotically approaches *n*, but it has been shown that this approach requires extremely large datasets [47].

Alternatively, many approaches utilize Bayesian inference to investigate this problem. Tang, Sindler, and Shirven assign a prior uniform distribution to *n* and beta distribution to *p* [49]. These beliefs are updates with the number of observed successes, *r,* resulting in estimates and 95% confidence intervals for *n*. Due to the beta parameters requiring assignment based on the statistician's confidence, this approach does not yield unique confidence intervals. Additionally, the example intervals shown by their approach are as much as 80 units wide (275% of estimated *n*) [49]. Similarly, Draper and Guttman utilize Bayesian methods to estimate a probability density

for p(n|x), that is the probability that n equals certain values given observed values x. This probability density can incorporate any prior beliefs held about p. The mode of the resulting probability density can be interpreted as n, but the authors also caution against inaccuracies that occur when no prior information is available concerning p [37].

For the problem examined in this research, the lack of prior information concerning p complicates Bayesian approaches and the small available datasets remove asymptotic approaches. My approach is applicable for problems where n is small. This research utilizes a combination of the method of moments and point-estimation calculations, but future work could examine the effect of altering the method of n and p estimation upon this approach's effectiveness.

2.2 Applied Method to Determine *n* and *p* in this Study

Utilizing a combination of the method of moments and point estimates, estimates of n and p can be obtained as follows: Rearranging Eq. (4) :

$$p = \frac{E[x]}{n} = \frac{\mu}{n} \tag{7}$$

This result can be substituted into Eq. (2), yielding

$$f(x|p) = \binom{n}{x} (\frac{\mu}{n})^x (1 - \frac{\mu}{n})^{n-x} \tag{8}$$

If μ can be calculated and $f(x|p)$ can be approximated from observational data, then Eq. (8) can be solved to reveal n for each number of successes x=1,2, … max(x). Max(x) is the maximum number of successes observed. These n values are then averaged and the resulting average n value is then substituted into Eq. (4) to yield an estimate for p at that location.

Confidence in this approach depends on if the value of n is small and p is large. Specifically, large n values and small p values will result in inaccuracies like the currently applied methods.

The first tool to minimize *n* is discretization. Recall that one is interested in the queue of total length *n*, observed *over some finite time period*. By limiting the finite time period observed, the number of potential users (*n*) observed is also appropriately limited. Further analysis of the system being evaluated must also be conducted to ensure that the system is amiable to this approach.

Once the dataset is discretized, the values of *n* and *p* for each location are determined as shown in the following pseudo-code. Each station is analyzed. The environmental conditions are those joint conditions under which the *r* datapoints are collected. For Case Study One environmental, see Table 1.

```
Begin
   For Each Station
         For each environmental condition:
               If dataset <cutoff
                     Eliminate those datapoints.
               End
         Use Eq. (7) to determine n.
         End
         Calculate n Estimate by taking a weighted average (weighted by r) of calculated n
         Use Eq. (6) to calculate p.
   End
End
```

If environmental conditions play a role in consumer behavior, only adequately represented conditions should be considered. This discrimination is for two reasons. First, conditions with an

insufficient number of observed hours could result in inaccuracies when estimating *n* and *p.* For example, a condition that only occurs ten times instead of one hundred times will result in wider fluctuations in estimate values for *n* and *p* for a sample change [47]. Secondly, outlier conditions (such as record hot or cold temperatures) should not be used to estimate future demand. In an ideal scenario, sufficient data would exist to estimate *n* and *p* for all possible conditions. If sufficient data existed, the resulting average *n* and *p* could be used to accurately estimate future hourly demand.

If all joint conditions are not considered, however, the resulting demand estimate should only be used as an ordinal measurement. E[x] provides only relative information about each location's demand. Coupling this approach to other methods could allow designers to assess the maximum demand anticipated and use those results to size the remainder of the system.

Once populated, the estimated *n* and *p* values can be used to predict *n* and *p* at new product or service locations. The calculated *n* and *p* estimates are plotted over the serviced area and fitted with thin-plate splines via Matlab. These splines are then used to create a density plot. By evaluating the density plot at each new location, *n* and *p* are determined for each new location. Eq. (4) is then utilized to determine E[x] at each new location, allowing designers to anticipate demand at new locations.

In summary, the algorithm approach for this RP demand estimation consists of the following steps.

1. For one PSS historical location, eliminate non-representative joint conditions. This remaining data is the PSS location historical data.
2. Evaluate the mean of the PSS location historical data. This is μ of the PSS location historical data.

3. PSS location historical data thus creates a Bin(n,p) with number of observed successes, *x* ranging from zero to max observed.
4. For each *x,* from zero to max observed, evaluate *f(x|p)* from the PSS location historical data. For example, for f(x==0|p) is the fraction of hours in the PSS location historical data where there were no checkouts.
5. Utilizing Eq. (8), calculate *n* for each *f(x|p)* for x from zero to maximum observed.
6. Utilizing Eq. (7), calculate *p* for each *f(x|p)* for x from zero to maximum observed.
7. Evaluate the average results of steps 5 and 6. This is the estimated *n* and *p* from one historical PSS locations. This is a weighted average, to ensure that the *f(x|p)* calculations based on more data points have a greater influence on the final estimated *n* and *p.*
8. Once steps 1-7 have been completed for all historical locations, fit a 3-D surface to both *n* and *p* for all historical locations. These are the *n* and *p* surfaces for the evaluated PSS. For this study, curve fitting was performed with the MATLAB command fit([lattitude, longitude,],n,'thinplateinterp','Normalize','on'). The thin-plate spline fits a surface to the *n* and *p* points by seeking to minimize a weighted sum of the surface error and the roughness of the surface [50]. The creation and use of this surface is one of the primary novel contributions of this work.
9. Once the *n* and *p* surfaces have been created, the expected *n* and *p* values can be retrieved for specific new PSS location coordinates within the service area.
10. To estimate the expected demand at the new locations, calculate E[x] from the expected *n* and *p* values with Eq. (4).

2.3 *n p* Estimation Validation and Verification

Several validation simulations were conducted to ensure the method of *n* and *p* estimation was accurate. First, a single simulation solved for *n* and *p* at various conditions with a large number of samples, *r* (*r*=5000). This simulation provides a best-case scenario estimate. As *r* increases, the observed data set distribution should accurately reflect the underlying distribution. Results shown in Figure 3 below. As expected, this *n* and *p* estimation approach performs best for small *n* values or large *p* values. Consistent with previous approaches, *n* is underestimated for data sets with large *n* values [47].

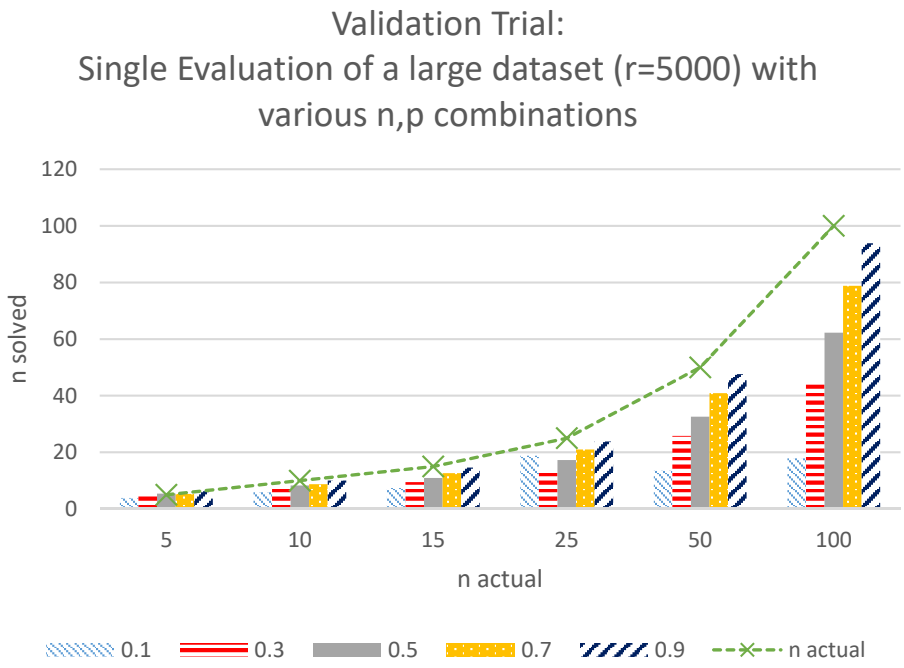


Figure 3: Validation Trial One

While the first validation simulation explored the accuracy of this estimation approach over a range of *n* and *p* for a single large dataset, the next validation simulation explored the impact of the value of *n* on the expected estimation accuracy for smaller datasets. Monte Carlo simulations were conducted with *r*=90 and *p*=.3, while different values of *n* were explored. Results are shown

in Figure 4. Nmonte is the result of the estimation algorithm while Nact is the actual *n* value for the tested population. The Error bars are one standard deviation (with negative error capped at 0 for graph readability).

This validation trial yielded several interesting results. First, one sees that the most accurate estimates occur at n=10 (97.7% of *n* actual) and n=5 (115.6% of *n* actual). The average *n* prediction performance did degrade from n=20 to n=25. These results highlight the importance of applying this estimation technique to datasets with appropriately small *n* values.

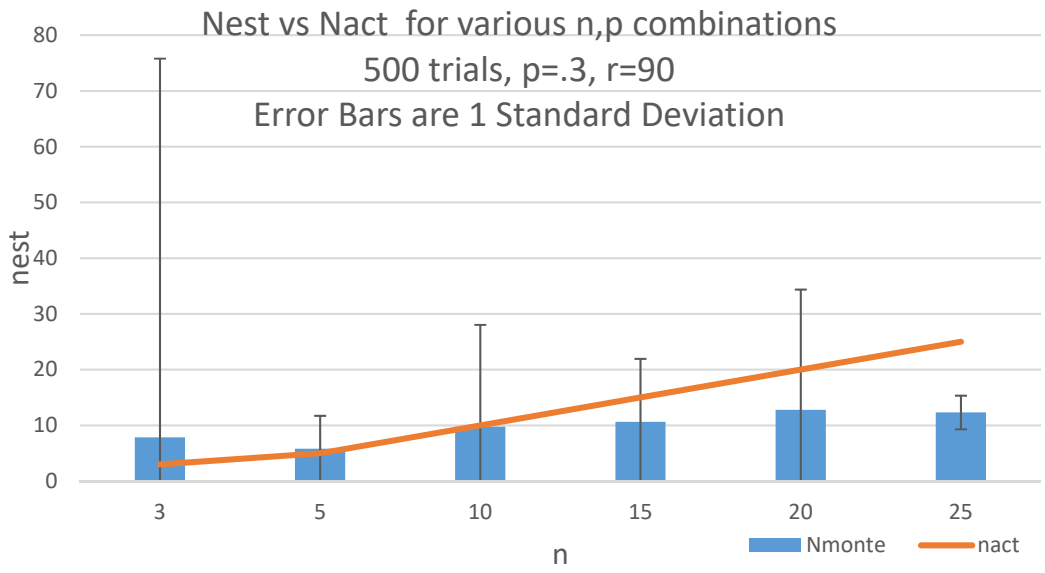


Figure 4: Validation Trial Two

The final validation simulation examined the impact of varying *r* on estimation accuracy. I began with the most accurate *n* value from the second validation simulation (*n*=10, *p*=.3). Then, Monte Carlo simulations (1000 trials) were conducted to assess the impact on estimation accuracy for r=20,40,60,80,100. The results are shown in Figure 5. The estimates varied from 8.01 to 8.45. This minor change indicates that the approach is not very susceptible to variations in *r* for the range 20-100. Of note, the second validation simulation did provide evidence that extremely large data

sets (large r) do provide value. The r=500 estimate from that trial was 9.77. Thus, although small variations in dataset size are not expected to influence the results of this study, this trial indicates that the best accuracy can be obtained when using larger datasets.

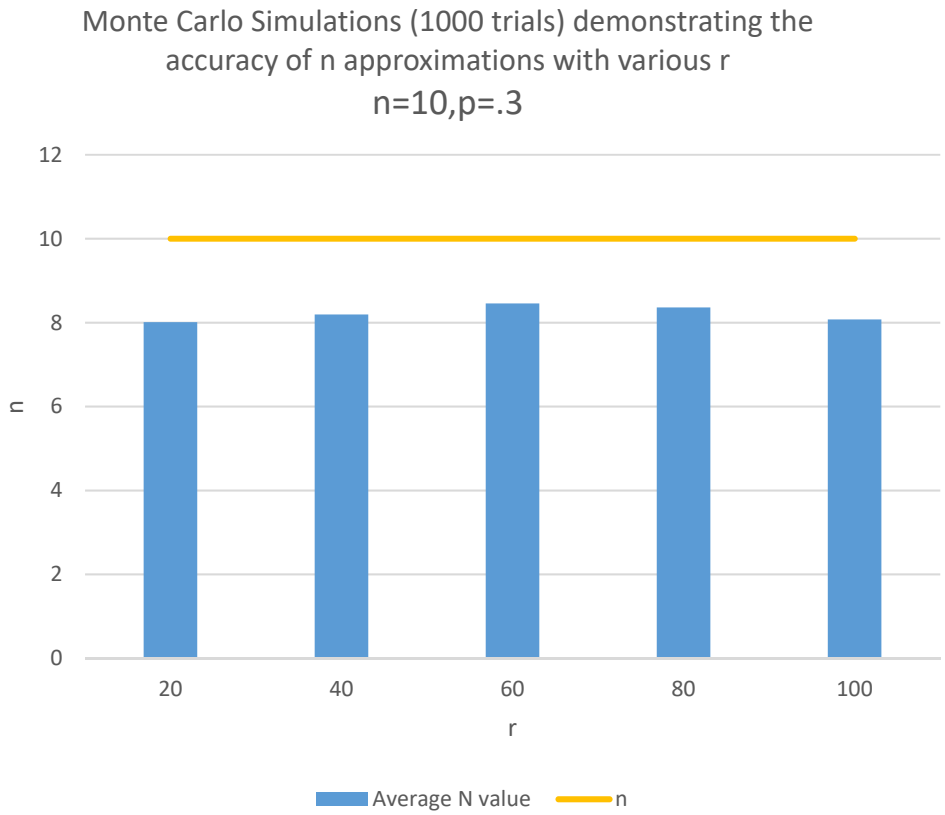


Figure 5: Validation Trial Three

Thus, these three validation simulations provide confidence and guidance for the proper application of the proposed *n* and *p* estimation approach. I examined the accuracy of the estimation approach over a large range of *n* and *p* values in the first simulation, confirming that this approach is best applied to systems with a relatively small range of expected *n* and *n* less than 25. Secondly, this approach is relatively robust with regards to variations in dataset size (*r*), only varying by 4.4% of *n* when increasing *r* by a factor of 5.

1.4 Summary

This chapter presented historic attempts to estimate *n* and *p* followed by the method chosen in this study to estimate *n* and *p* from historic user data. Finally, three simulations validated and provide application guidance for *n* and *p* estimation approach. Next, the estimation approach will be applied to two case studies. Chapter 3 continues the investigation by describing and testing a methodology to apply the *n* and *p* estimation approach to an increase in Product Service System density in Case Study One.

CHAPTER 3: Case Study One 2015 DIVVY Expansion (Within the Boundaries) Background and Methodology

The 2015 Chicago BSS expansion provides an excellent case study to test the efficiency of using *n* and *p* estimation to provide Product Service System Design insights. The Divvy BSS was launched in June 2013 growing from 75 to 299 stations by 2014. The average size of the 299 stations existing in 2014 was 17.4 bicycle docks with a maximum of 43 and a minimum of 11. Ridership was automatically recorded by the time the bikes were checked out from each dock. This data was autonomously recorded by Divvy and provided via https://www.divvybikes.com/system-data. A current interactive Divvy map is available at https://member.divvybikes.com/stations.

At the beginning of 2015, Divvy added 176 additional stations [51]. Crowdsourcing, input from elected officials, surveys, and community events were used to help plan the expansion [52]. There were no restrictions on the possible locations of the new docks, however this investigation revealed that the operators placed the docks density consistent with the Institute for Transportation and Development Policy (ITDP) guidelines of 10-15 docks per km^2.

Riders are divided into subscribers (who purchase annual passes) and customers. In 2014 1,663,394 subscribers took rides, while in 2015 that number increased to 1,990,310 of which 263,103 were from the new stations (Figure 6). Riders are divided into subscribers (who purchase annual passes) and customers. By utilizing the 2014 subscriber ridership data to create the *n* and *p*

surfaces for Chicago, the demand estimation approach proposed in this thesis can be tested by predicting the demand for the 176 stations installed in 2015.

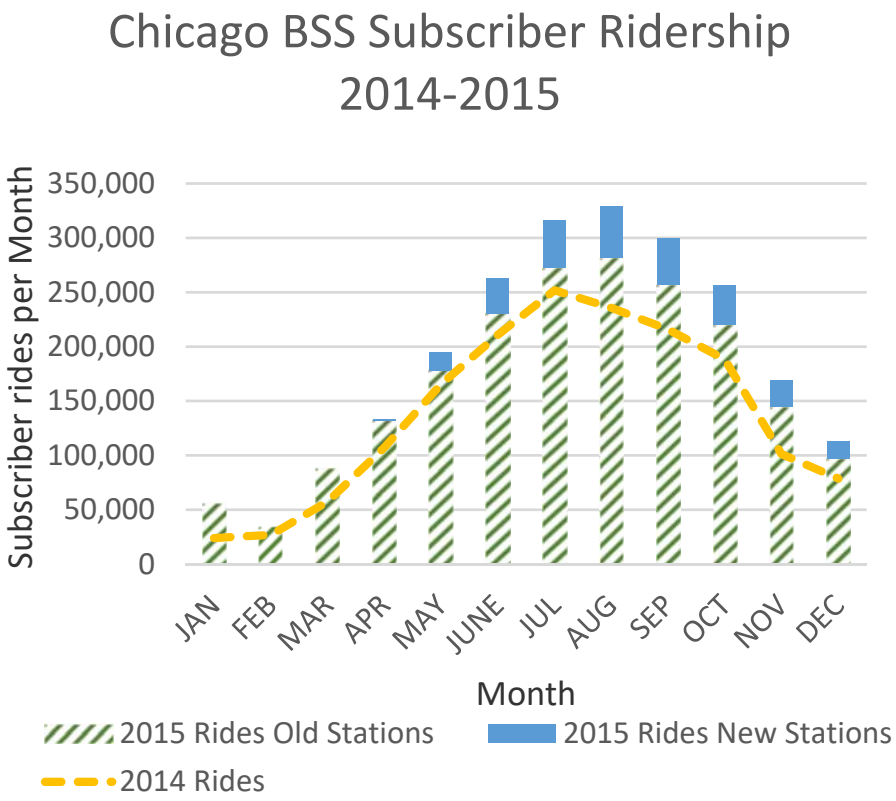


Figure 6: 2014-2015 Chicago Bikeshare Ridership

3.1 Background

This section will present necessary background information to fully understand the methodology employed in Case Study One. First, to increase understanding of the role demand estimation plays in the BSS industry, current station sizing methods are presented and described. Then, current BSS modeling work is presented to both highlight the difference between this study and previous efforts and provide evidence for methodological choices.

3.1.1 Case Study: Sizing New Stations

There are two primary sources of guidance for BSS expansion. The National Association of City Transportation Officials (NACTO) published a *Bike Share Station Siting Guide* in 2016, while *The Bike-Share Planning Guide* was produced by the Institute for Transportation and Development Policy (ITDP) [10,43]. These two guides include data or collaboration from a variety of major BSSs including New York City, Chicago, Washington DC, and Paris.

Although NACTO acknowledges that BSS docks can be configured for a variety of sizes, its guide focuses on new BSS dock locations rather than station sizing [43]. ITDP encourages operators to size the station using three parameters: bike station density, bike density, and bikes per station. The appropriate bikes per station are driven by local demand. ITDP recommends the operators conduct surveys, evaluate points of interest, crowdsource location ideas, use local knowledge, or hold community workshops to determine station sizing [10]. These methods could be susceptible to biases such as small study effects, selection bias, or response bias. No quantitative method is provided in these guides to determine new station sizing.

The global average BSS station sizing variance reflects this lack of standardized guidance. An analysis of 38 global systems revealed only a weak correlation (R^2=.1) between average docking station size and system size. Average BSS dock sizes varied from 12-34 [53].

3.1.2 Case Study: Previous Algorithmic Work

Research in BSS modeling and development of tools for BSS operators has grown with BSS popularity. Although a full literature review is beyond the scope of this investigation, a brief overview will present role the approach presented in this thesis has in relation to ongoing efforts.

Much of BSS algorithm research is driven by the problem of rebalancing [25,26,28,51,54]. Rebalancing refers to the process where vehicles and personnel relocate bikes to compensate for

asymmetric demand patterns. One study revealed that bike sharing increased the overall motor vehicle usage when the effect of bike rebalancing was considered in London [54]. Due to the negative environmental and operational impacts of rebalancing, BSS researches have been motivated to develop algorithms that assist in predicting existing station demand. Unlike the research in this thesis, which focuses on design phase demand estimation for new BSS stations, these algorithms have the goal of improving BSS operations by predicting usage in close to real time and anticipating atypical demand transients for existing BSS stations.

Some work approaches the problem of rebalancing by focusing on predicting individual trip destinations to allow operators to anticipate asymmetric flow patterns [51]. A study of the Divvy BSS utilized information about the departure time, user characteristics, and historic station travel pairs to predict destination station and trip duration, achieving an impressive 87% destination accuracy [51].

Many studies focus on predicting station demand using approaches such as computer generated decision trees, probabilistic topic models, naïve Bayesian networks, and integer programming of demand heuristics [25–28]. Demand predictions for existing stations can be used to anticipate when rebalancing will be necessary. In this approach, researchers model bike station usage directly, rather than modeling individual bicycle movement [27].

Specifically relevant to this thesis is previous efforts by Rudloff and Lackner [55]. They showed that a combination of negative binomial, Poisson, or hurdle models can be effectively utilized to predict hourly station demand. Their approach predicted individual station hourly demand using generalized linear count models derived from historical data. The approach in this thesis builds upon their success utilizing count models to describe BSSs with three clear differences. First, I utilize existing station data to forecast demand at new stations in new locations

that have no historical data; Rudloff and Lackner predicted ridership at existing stations and existing locations. Secondly, Rudloff and Lackner's approach utilized hurdle, negative binomial and Poisson models. These models were chosen to focus on maintaining bike availability after install and design, while the approach presented in this thesis utilizes the binomial distribution to focus on design and sizing of the station before install. Focusing on determining n and p provides additional information to the decision maker. The distributions examined by Rudloff and Lackner, however, aim to identify how long before an event occurs (such as the station becomes empty). Finally, the furthest prediction look ahead Rudloff and Lackener reported was two weeks, while this effort predicts ridership over a year.

To the author's knowledge, there is limited work on using an algorithmic approach to determine new BSS station sizing; many efforts stop analysis at categorizing and clustering existing stations [27,56,57]. One study applied a regression approach to determine the effect of 16 factors around each station on utilization. Although these results could be used to plan new station location and sizing, the authors caution that their results reflect BSS startup transients and may not be applicable to other systems. Additionally, their approach examined existing BSS, rather than performing predictions for new BSS [57]. A study of the Velib system in Paris analyzed one month of data to sort stations into usage profile clusters [56]. These results were then compared to four environmental factors to examine the relationship between the usage patterns and the environmental factors. Finally, in Barcelona various utilization scores were calculated to describe daily usage patterns. An algorithm was utilized to hierarchically cluster the stations utilizing only these scores, omitting geographic information. The clustering revealed that the stations were geographically clustered [27].

This thesis builds upon the previous clustering efforts. These studies show that geographic or usage characteristic clustering of existing stations is an effective framework to examine BSS. This validates a key assumption for the approach presented in this thesis: that geographic locations can provide insight into demand patterns. This thesis expands upon previous clustering work, providing a framework to apply the underlying BSS demand at existing stations to predict ridership at new stations.

Additionally, previous efforts prediction algorithms were effective with a limited look ahead time, limiting their effectiveness as tools for long-term BSS planning [25,27]. These studies predict existing station occupancy looking ahead ten minutes to a week into the future to assist in day-to-day BSS operations. Contrasting, the approach in this thesis seeks to estimate the underlying user demand characteristics to support design decisions. This thesis seeks to predict expected steady state demand, not product adoption dynamics or specific demand transients around the steady state. Thus, this approach is appropriate for look ahead times long enough for planning, but assumes that user demographics and environmental characteristics do not shift significantly. The approach in this thesis could be useful for planning for 8 months – 5 years, but follow-on research is necessary to identify the appropriate prediction period.

3.2 Methodology

Applying the *n* and *p* estimation approach presented in Chapter 1 requires several steps. First, the BSS must be evaluated to verify that the characteristics of that system match binomial Distribution Characteristics. Next, *n* and *p* must be defined and practical estimation problems for the BSS application must be overcome. Third, the general algorithm described in Chapter 2 must be adapted to this case study. Finally, metrics to determine the efficiency of this approach must be presented and justified.

3.2.1 Interpreting BSS as a Binomial Distribution

For a BSS system, *n* describes the station's population size, and *p* describes the population's BSS affinity. *n* and *p* are independent, fixed, and unknown [37,39]. As the station is observed over multiple comparable finite periods of time, this creates a dataset, such that X_i (i= 1, 2, … r) [40]. Where *X* is independently and identically distributed from Bin(n,p).

It seems evident that between stations the number of individuals considering riding (*n*) could vary depending on location. Literature into why people ride BSS provides insight into the possible sources of inter-station variation of *p.* For example, many users choose to use BSS due to convenience over other options [42,45]. Logically, this preference could vary throughout the BSS area as other forms of travel infrastructure vary.

3.2.2 Chicago BSS *n* and *p*

Due to historical efforts to estimate *n* and *p* having difficulty when *n* is large and *p* is small, I attempted to minimize *n* four ways. First, define *n* to be all people who are both subscribers and are currently making the choice to travel. *n* includes individuals who did not chose to utilize DIVVY by selecting another mode of transportation (including private bicycle or car), but were subscribers and could have chosen DIVVY to travel at that time. Secondly, the finite time interval examined is one hour. Third, to ensure the same population was considered during each hour, the data was discretized and sorted as shown in Table 1**.** The wet-bulb temperature was utilized to allow implicit incorporation of humidity effects.

2014 and 2015 rider data are available from Divvy, while hourly weather data were obtained from NOAA. NOAA data was collected at Chicago O'Hare Airport, approximately 22.5 km (14 miles) from the center of the BSS. The following terminology will be used to discuss the BSS data.

A joint condition refers to the joint environmental conditions which exist (temperature, hour, weekday, and precipitation). A data point is the number of rides observed in an hour.

Utilizing all data to estimate n and p was not implemented for this case study. It is probable that some joint conditions have very low n such that they might bias sizing to be too low. For example, it is possible that one person rode at 3am in the rain on a Tuesday. There are likely a number of joint environmental conditions (i.e. night-time, cold, rainy) that would artificially lower the estimated n and p. Joint conditions with less than 90 observed hours were eliminated to prevent non-representative joint conditions from skewing the overall results. 90 hours per joint condition was selected to prevent erratic estimation fluctuations for n and p due to a small sample change by maximizing rp^2 [48]. 90 is approximately 1% of the hours recorded in 2014.

Only the hours of 5,6,7,8,11,12,13,15,16,17,18, and 19 were considered (correlating with morning, lunch, and evening rush hours). These constraints resulted in 15 of 384 observed joint conditions being utilized and 30.9% of the observed hours being utilized. Although 30.9% of the hours initially appears low, recall that only utilizing the 12 peak transit hours limits the available data to 50% of the recorded hours. The reduction from 50% to 30.9% of the recorded hours removes the uncharacteristic environmental conditions shown in Table 1.

Table 1: Discretization Utilized for Case Study One

Category	Number of Bins	Bin Edges
Time	24	Each Hour
Wet Bulb Temp	4	-20F-5F, 5F-30F, 30F-55F, 55F-80F
Precipitation	4	0-.5",.5"-1", 1"-1.5",1.5"-2"
Weekday	2	Weekday or Weekend

Fourth, as of March 2014 Divvy had 14,000 subscribers [58]. A simplistic calculation shows that 14,000 subscribers spread over 298 stations and 24 hours yields 1.95 subscribers/station-hour (assuming no repeat riders in a day). As a more complex analysis, shown in Figure 7, simulations were performed to assess the maximum expected mode if all subscribers decided to ride simultaneously. The first simulation assumed all stations were equally preferable and revealed 97% of the time the mode fell below 80.

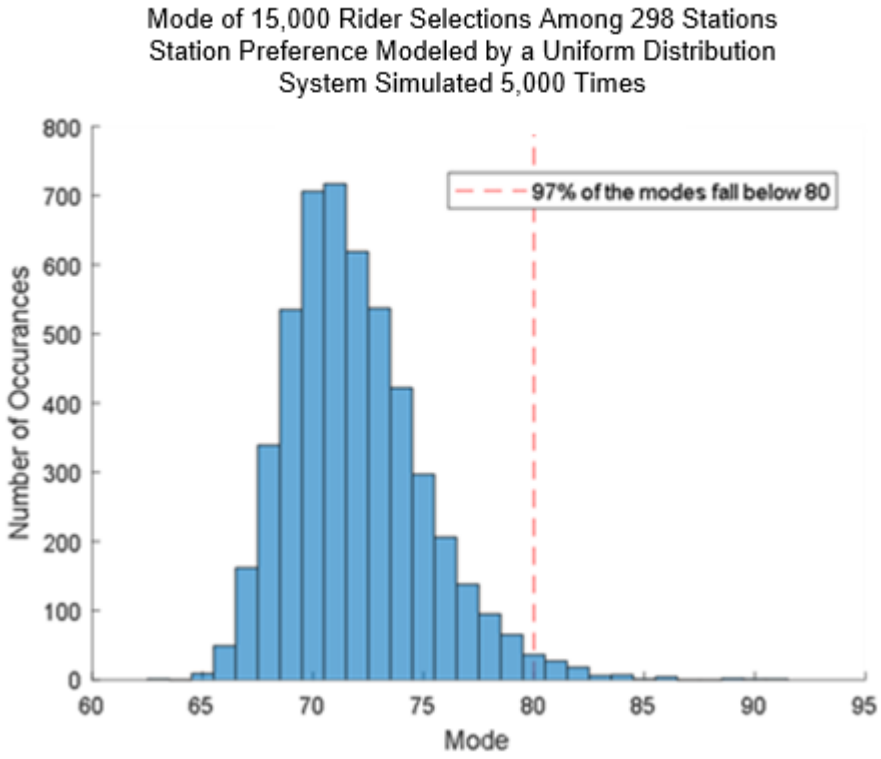


Figure 7: Simulated Rider Choices with Uniform Preference

Figure 8 shows a second simulation that assumed the rider station preference matched actual user history. The rider choices of the thirty busiest hours of 2014 were used to model user station preference, yielding a mode of 405.

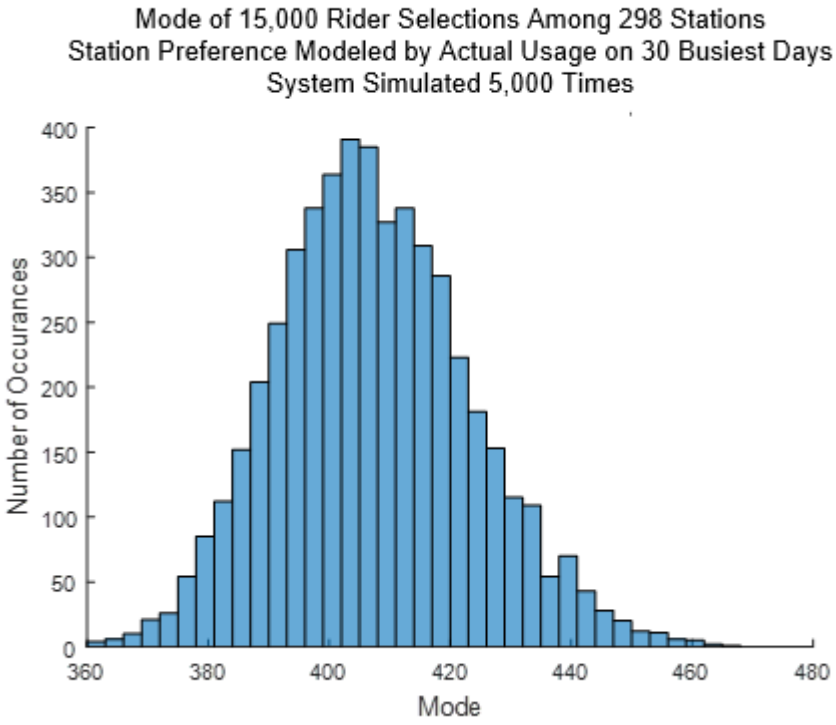


Figure 8: Simulated Rider Choices with Historical Data Preferences

Although 405 and 80 riders per station are larger than desired for this approach, the busiest hour of 2014 only had 1,468 riders, not 15,000. Scaling these simulations to assume only 1,500 riders instead of 15,000 in the same hour yields maximum expected demand of 8 and 40.5, sufficiently low for estimation.

Finally, recall that for MLE this erratic estimator behavior is due to the fact that when σ^2/μ^2 is near one, the likelihood function is essentially level [39]. For the 2014 Chicago BSS 14.7% of the stations analyzed had σ^2/μ^2 between 0.5 and 1.5.

3.2.4 Case Study Algorithmic Specifics

The general approach taken to estimate PSS *n* and *p* is explained in Chapter Two Section 2.2. Case-study specific decisions are described in this section.

n and *p* were evaluated in accordance with the prior pseudo-code where "*environmental conditions"* were "*temperature, time, weekday, and rain scenario".* One station (283) was unable to yield an estimate for *n* or *p*. This station was omitted from the analysis.

Due to the fact that the environmental conditions used to estimate *n* and *p* are potentially different than future demand situations, the resulting demand estimate from this approach should only be used as an ordinal measurement. E[x] provides relative information about station demand. Coupling this approach to traditional methods (crowd-sourcing, demand surveys) could allow BSS operators to set the size of the largest BSS station and utilize these results to order the remaining stations.

Figure 9: Boundaries for Case Study One

Of the 176 stations added in 2015, the 46 stations lying within the boundaries of the 2014 stations were evaluated. The distributed PSS demand estimation approach presented in Case Study Two depends upon estimating *n* and *p* at the current station or PSS locations and using that data to estimate *n* and *p* at new PSS locations. Thus, the appropriate scope of application is limited to new stations geographically located inside the boundaries of the data used to derive the *n* and *p* surfaces. This approach is applicable to a PSS expansion that increases PSS density, not an expansion that increases the area serviced by the PSS.

2014 stations and stations added during 2015 expansion are shown in Figure 9**.** The small dots represent the 299 DIVVY stations already installed by 2014. These stations provide the data used to estimate the *n* and *p* surfaces. The larger circles are the 46 added stations evaluated by this case study. The symbol in the center of each circle indicates the number of docks in each new station. The new dock sizes are generally largest closer to the center of the Chicago Business District.

When creating the *n* and *p* density plots, all values within a 0.4 km (1/4 mile) edge box were averaged. 0.4 km (1/4 miles) was utilized to be consistent with current "last mile literature." Most individuals will walk between 0.4 km and 0.8 km (¼ and ½ mile) to reach transportation [59].

3.2.5 Comparison of Algorithm and Operator Choices

The implemented station sizing ranking and algorithmically recommended rankings were compared to the ideal ranking with an interrater agreement test. The ideal ranking was determined by ranking the 2015 station utilization from highest to lowest. Because the 2015 new stations entered service between April and July, only ridership from August to December was utilized to make the ideal ranking.

Spearman's rho (ρ), or rank correlation, allows for comparison of rankings for non-parametric data, as defined by Eq. (9) [60]**.**

$$\rho = 1 - \frac{6\sum d_{i\,i}^2}{s^3 - s} \tag{9}$$

Where d_i is the difference between each pair of rankings and s is the number of stations ranked (usually 46 for this study).

To resolve differences in Spearman's rho due instances where the operator ranking had rank ties and the algorithm did not, a correction factor was utilized for ties. Ties exist in the rankings of the sizes implemented by the operators and recommended by the algorithm. For operator rankings this is because only six distinct station configurations were used for the 46 stations. For algorithm rankings, local minimums in the n and p curves can result in negative E[x] values. These locations were analyzed as having an E[x] of zero. It is necessary to adjust the value of Spearman's rho to ensure that there is a fair comparison between Algorithm results with few ranking ties and Operator rankings with many ties. Using Eq. (10) and Eq. (11) adjusts the value of Spearman's rho for ties to allow equivalent comparisons [60].

$$\rho = 1 - \frac{6(\sum d_i - \sum t_x)}{s^3 - s} \tag{10}$$

Where $\sum t_x$ is

$$\sum t_x = \frac{\sum(t_i^3 - t_i)}{12} \tag{11}$$

and t_i is the number of rankling ties.

3.3 Summary

This chapter presented the necessary background and methods utilized in Case Study One. Case Study One uses binomial distribution parameter estimation to estimate demand for a Product Service System expansion. The 2015 DIVVY Expansion provides a case study to test this approach. Next, in Chapter Four, the results of Case Study One will be evaluated and compared to the choices made by the DIVVY operators to validate the effectiveness of the approach in this thesis.

CHAPTER 4:

Case Study One 2015 DIVVY Expansion (Increase in System Density) Results

While Chapters 1-3 presented the need for improvements Product Service System (PSS) demand estimation, a frame work for estimating *n* and *p,* and the methods for applying that framework to an increase in PSS Density, this chapter presents the results from the first of two case studies, discusses limitations of this approach, and motivates Case Study Two.

4.1 Results

Case Study One resulted in two types of results. First, the actual generated *n* and *p* surfaces themselves were analyzed for insights. Secondly, the accuracy of the algorithmic predictions were compared to the implemented operator ordering.

4.1.1 Divvy BSS Expansion Surfaces

Once the data were evaluated, the *n* and *p* surfaces were calculated. Figure 10 shows both a 3-D and contour plot of the *n* and *p* surfaces. The surfaces cover approximately 337 square kilometers (130 square miles). The *n* surface is generally flat, but is dominated by seven peaks with estimated *n* of greater than 18 (stations 35, 75, 141, 143, 192, 238, 281). The peaks may be due to high demand areas such as major employment areas or parks. For example, station 35 has the largest *n* of 34.2 and is near the Navy Pier and Chicago Children's Museum. Stations 143 and 144 are near Lincoln Park. Stations 75 and 192 are near Chicago's Union Station. The *p* surface, however, is much rougher and is characterized by a few local peaks, indicating that people in those areas are more likely to utilize BSS than those on the edges of the BSS.

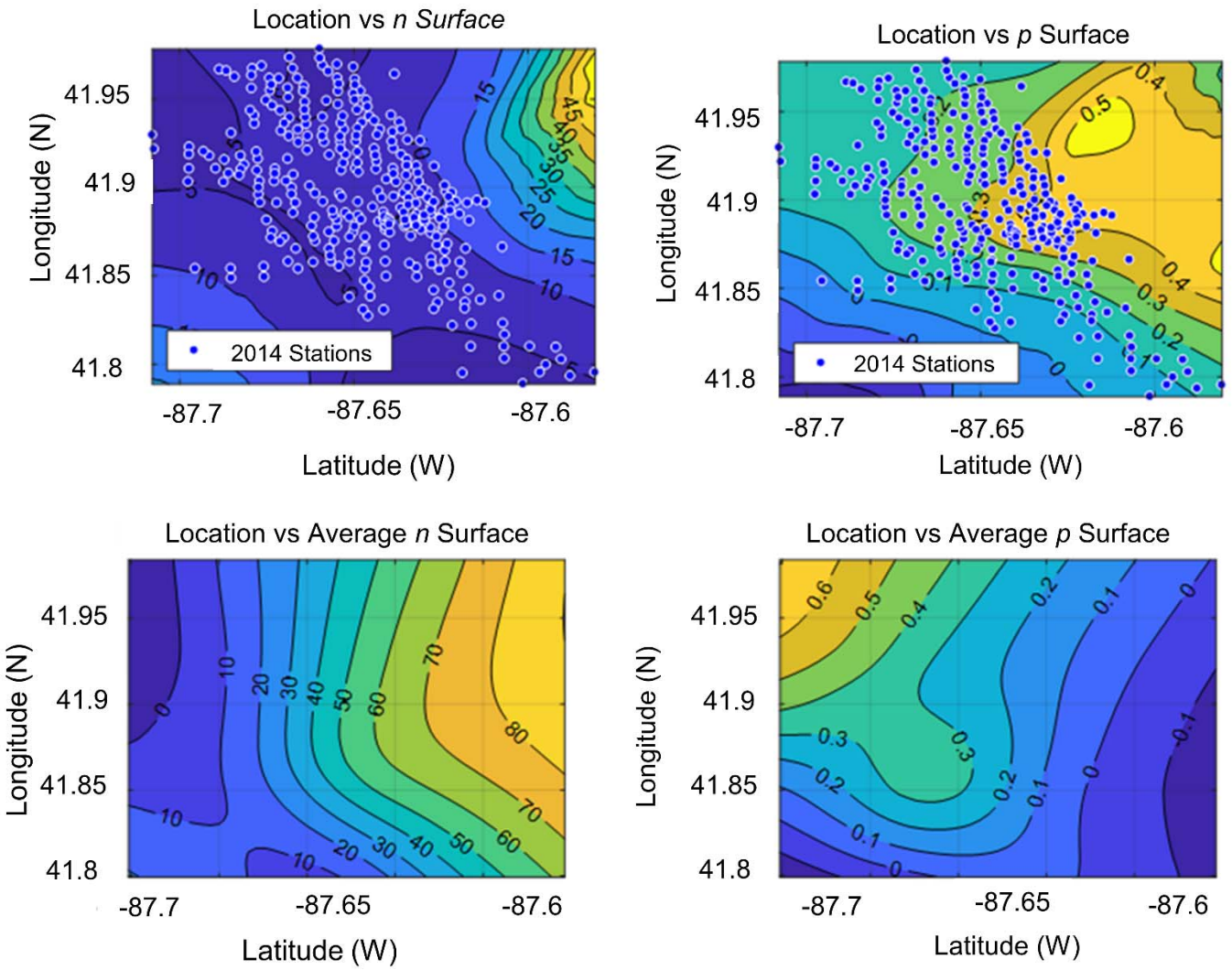


Figure 10: n and p Surfaces

Figure 10 containts contour plots of the average *n* and *p* surfaces, created by averaging all values over a ¼ mile radius. The average *p* surface is dominated by a peak centered roughly over the central business district of downtown Chicago.

Of note, the resulting surfaces depend upon the method of *n* and *p* estimation approach utilized. Future work should examine the effect on this approach of altering the *n* and *p* estimation approach. This dataset was also examined via the method of moments, but only yielded negative values for *p.* This is because the BSS dataset has variance greater than observed mean, rendering the method of moments ineffective.

4.1.2 Comparison to Operator Selected Ordering

Table 2 includes the ideal ordering (ordered by observed 2015 usage), algorithm ordering, operator selected ordering, calculated *n* and *p,* the resulting E[x], and the observed average hourly usage (μall) from August to December 2015. The average hourly usage (μpeak) and standard deviation (σpeak) during the joint conditions used to estimate *n* and *p* are also provided. Interrater agreement was calculated using Spearman's rho. The operator selected ordering was determined by ranking the stations added in 2015 from largest to smallest. The operator ordering showed a moderate correlation with the ideal ordering (rho=.60, stations=46, p<.01). The algorithm ordering outperformed the operator ordering, showing a very strong correlation with the ideal ordering (rho=.83, stations=46, p<.01). This difference included correcting for ties in the operators ranking by using Eq. (10) and (11). The improvement in spearman's rho indicates the potential potency in this work's RP demand estimation approach for distributed PSS.

To provide a qualitative assessment of algorithm accuracy, instances where algorithm or operator ordering varied from ideal ordering by more than ten places (1/4 of the sample size) are highlighted in gray. This occurred eight times in the algorithm ranking and eighteen times for the

operator ranking. Instances where predicted ordering was within four (1/10 of the sample size) of the ideal ordering are shown in bold. This occurred eighteen times for the algorithm ranking and twelve times for operator rankings.

Utilizing 2014 average ridership was also effective at predicting 2015 ridership for the 2015 rides at the old stations (Pearson's Correlation of 0.91 between 2014 μ and 2015 μ). Finally, this analysis was repeated only utilizing observed average hourly utilization for predication rather than *n* and *p*. This analysis was done to determine if utilizing *n* and *p* rather than μ improved prediction accuracy. Pearson's Correlation between predicted E[x] dropped to 0.3473. This is less than the 0.3813 correlation between the traditional method used by the BSS operators and μ. Spearman's rho also degraded negligibly, from 0.83 (*n* and *p*) to 0.80 (E[x]). Therefore, utilizing an E[x] surface did provide superior Spearman's Rho to traditional methods, but slightly inferior results to using *n* and *p* surfaces. The degraded performance of the E[x] surface supports the assertion that utilizing *n* and *p* provides additional information to the designer than hourly utilization alone.

A major performance difference is that the operator ordering is more accurate for the highest performing stations. This result indicates that areas of highest interest are easier to distinguish from areas of low or medium interest. The traditional methods of surveys and local knowledge appear to be more adept at identifying the areas of highest demand. These areas also provide the most interest in BSS expansion. When Spearman's rho is recalculated without including what the BSS operator identified as the top four performing stations (top 10% of sample size), the algorithm still yielded a strong correlation to the ideal ordering (rho=.79, stations=42, $p<.01$), while the operator ordering revealed only a moderate correlation (rho=.48 stations=42, $p<.01$). These results indicate that the most advantageous approach might be for BSS operators to identify the highest priority

Table 2: 2015 DIVVY Expansion Prediction Results

Station	Ideal Order	Algorithm Order	Operator Order	*n*	*p*	E[x]	μ_{peak}	σ_{peak}	μ_{all}
133	1	14	**5**	7.9	0.44	3.5	6.9	5.9	3.3
3	2	**5**	**2**	11	0.40	4.6	4.2	6.1	3.2
18	3	12	15	8.5	0.43	3.6	3.9	4.2	1.5
38	4	10	**5**	8.4	0.46	3.9	3.6	4.2	1.7
4	5	23	**7**	5.4	0.39	2.1	3.1	4.0	2.2
142	6	**2**	15	16	0.48	7.7	3.1	2.5	1.5
107	7	21	**7**	6.6	0.35	2.3	3.0	3.0	1.4
89	8	16	2	8.4	0.33	2.8	2.8	3.4	1.2
39	9	**7**	**7**	10	0.40	4.1	2.8	3.5	1.4
6	10	1	2	26	0.35	9.3	2.7	3.7	2.0
145	11	4	15	9.6	0.55	5.3	2.6	2.7	1.5
161	12	9	7	7.8	0.51	4.0	2.5	2.4	1.4
125	13	8	28	8.5	0.48	4.1	2.5	2.7	1.2
41	14	19	**15**	8.3	0.32	2.6	2.5	2.3	1.4
7	15	3	**15**	18	0.42	7.5	2.5	2.9	1.4
2	16	**13**	1	10	0.35	3.5	2.5	3.9	1.8
96	17	**18**	7	6.6	0.41	2.7	2.2	2.2	1.1
182	18	**17**	7	7.1	0.39	2.8	2.1	2.5	1.2
364	19	**20**	28	6.4	0.36	2.3	2.0	2.9	0.91
359	20	**22**	15	6.3	0.33	2.1	2.0	2.4	0.95
172	21	11	45	7.5	0.50	3.8	1.7	1.9	1.11
383	22	30	7	3.7	0.25	0.92	1.7	2.2	0.69
40	23	15	28	8.7	0.40	3.5	1.6	2.5	0.68
180	24	6	28	8.5	0.53	4.5	1.4	1.9	0.89
417	25	36	15	4.9	0.12	0.60	1.0	1.2	0.67
374	26	34	**28**	4.5	0.17	0.75	0.91	1.1	0.59
370	27	**26**	**28**	6.0	0.28	1.7	0.86	1.4	0.45
103	28	**25**	**28**	7.1	0.25	1.8	0.85	1.2	0.40
465	29	35	7	4.5	0.15	0.7	0.82	1.1	0.56
507	30	**29**	15	4.4	0.21	0.93	0.75	1.2	0.56
365	31	24	**28**	5.3	0.36	1.9	0.58	0.95	0.28
372	32	40	15	5.8	0.07	0.40	0.54	0.91	0.34
402	33	42	28	5.3	0.03	0.18	0.53	0.86	0.33
403	34	41	28	4.9	0.04	0.21	0.49	0.87	0.31
502	35	28	15	3.5	0.31	1.1	0.46	0.86	0.26
504	36	**32**	28	4.3	0.19	0.84	0.44	0.73	0.32
501	37	27	28	3.1	0.38	1.2	0.40	0.78	0.31
505	38	33	28	7.0	0.11	0.80	0.40	0.71	0.20
414	39	**39**	28	5.5	0.08	0.44	0.33	0.65	0.23
401	40	**43**	45	5.5	0.03	0.17	0.28	0.67	0.16
413	41	**37**	28	5.1	0.10	0.53	0.19	0.50	0.10
366	42	**46**	28	6.8	0.01	0.07	0.13	0.38	0.09
410	43	**45**	15	3.1	0.03	0.08	0.10	0.34	0.05
416	44	38	28	4.5	0.11	0.51	0.09	0.33	0.06
406	45	31	15	6.0	0.15	0.87	0.09	0.33	0.05
407	46	**44**	15	2.3	0.06	0.14	0.04	0.22	0.02

Spearman's rho		
All	**.83**	**0.59**
Bottom 90%	**.79**	**0.49**

Note: Bold numbers are within 10% of ideal order, gray boxes are greater than 25% from the ideal order.

areas utilizing a low-cost, non-deterministic method such as local knowledge, followed by algorithmic assessment.

The predicted hourly utilization can also provide insights into the expected average hourly utilization. Figure 11 plots a linear regression between predicted E[x] and observed μ with a R^2 value of 0.43. Station capacity and observed μ are also plotted, with a slightly lower R^2 value of 0.38, indicating that algorithm results provides a slightly better insight into expected utilization of the new stations. E[x] was determined with the algorithm by using Eq. (4). Figure 11 also provides a visual indication of the range of actual ridership for the same implemented station size. For example, of the 13 stations installed with a capacity of 15, average peak hourly ridership ranged from 0.02 to 4.31 rides/hour. This correlation is consistent with previous studies (Table 3), which have resulted in R^2 values of 0.43, 0.381, and 0.476. Some studies have reported correlation of greater than 0.8, however these studies utilized monthly rentals or total demand, rather than *n* and *p* as done in this study.

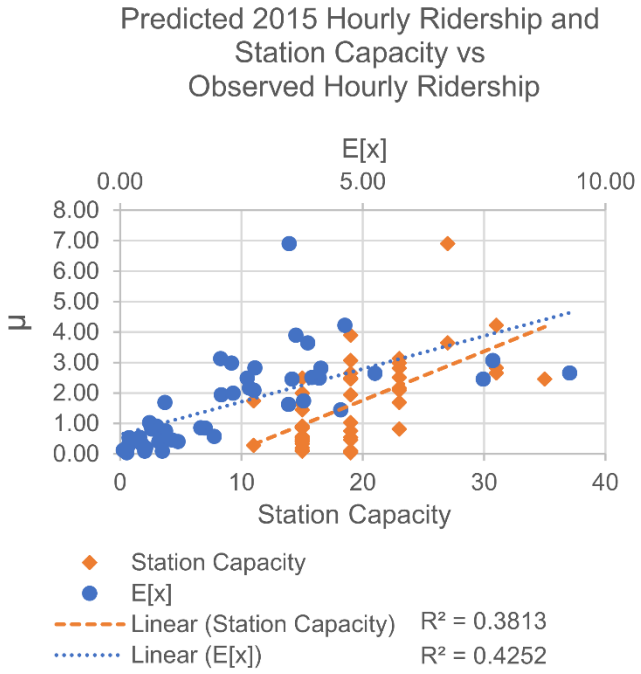


Figure 11: Predicted E[x] and Selected Station Capacity vs Observed Hourly Utilization(μ)

Although the current methods employed by operators were effective at identifying the most popular stations, the algorithmic approach was superior at identifying the demand for the medium to low demand stations, indicating a promising approach of utilizing these approaches in tandem.

4.1.3 Sustainability Impact of Algorithm Implementation

While rank data provides insight into the efficiency of this approach, it is also necessary to estimate the real-world benefit of applying this approach instead of the approach utilized by the actual BSS operators. For the 46 stations examined in this case study, a total of 890 docks were added. If one assumes 890 as the "new dock budget", one can estimate the algorithmic placement with the following formula which ratios the 890 available docks by each stations' expected utilization. This approach is repeated using μ to identify the ideal placement.

$$StationSize_i = 890 * \frac{E_i[x]}{\sum E[x]} \tag{12}$$

When comparing the ideal placement to the recommended algorithm and actual station placement the algorithm places 482 docks in optimal locations (54%), while the operators only placed 388 docks in optimum locations (44%) of the stations. This means that the algorithm placed 94 docks (10% of those added), equivalent to approximately 5,640 pounds of steel, in more desirable locations than the operators did. As an added benefit, this approach identified three stations with a recommended capacity of only one dock (stations 407, 410, 366). The ideal station size for these stations were 1, 0, and 1 dock. The operator outfitted these stations with 19,19, and 15 docks. This indicates that this capacity estimation approach may provide a method to check if proposed station locations are viable.

4.1.4 Sensitivity Analysis

Consistent with historical approaches to simultaneously estimate n and p, validation simulations that tested this work's n and p estimation approach against simulated datasets also

showed that this approach to estimate *n* is susceptible to both underestimation and fluctuations [47]. A Monte Carlo Simulation was conducted to estimate the expected Spearman's rho of this work's approach when considering inaccuracies due to *n* or *p* estimation. The *n* value for each station was selected from a normal distribution based on the mean and standard deviation of estimates of every observed hour (step 7 of the procedural summary). The results are shown in Figure 12. Although the mode of the Monte Carlo analysis is less than the Spearman's rho obtained during the case study, it is greater than the value obtained by the operators. This indicates that one can reasonably expect this approach to produce superior overall results to currently employed methods.

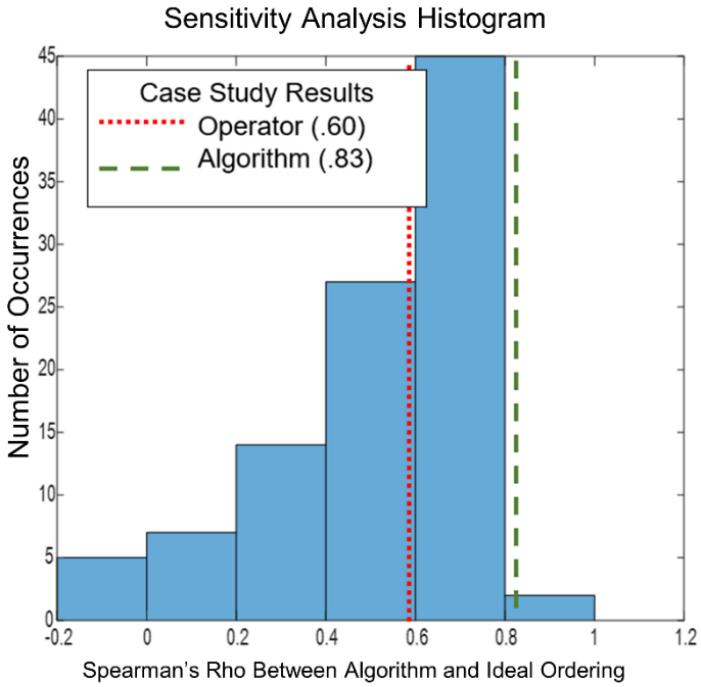


Figure 12: Histogram of Spearman's rho for 100 Monte Carlo Trials

4.2 Limitations and Future Work

This thesis proposed a RP demand estimation method to assist designers planning product or service expansion. By providing a design tool to anticipate user demand, this approach can provide critical design data for infrastructure expansion of PSS and other distributed systems. This thesis expands upon the current research thrust of in-situ demand transient estimation of installed PSSs, an operational vice design problem. Additionally, this work provides a possible tool to remedy the gap that currently employed SP, SI, and RP demand-estimation methods may not be suited to PSS. This is accomplished with a method to utilize user data to estimate demand in new situations based on binomial parameter estimation. The binomial parameter estimation is then used to create two surfaces that can be used to inform designer decisions planning an increase in PSS density.

This approach was validated with a case study that analyzed the 2015 Chicago BSS expansion. This approach yields results significantly more accurate than current demand estimation approaches. Traditional methods are effective at identifying the top performing stations, indicating that effective use of this tool is as an aid to designers after they have determined peak demand locations. Additionally, the case study builds upon previous algorithmic approaches that examine existing stations, demonstrating that current user data can be effectively used to estimate demand at future station locations.

The analysis of the BSS case study provides several insights about this thesis' RP demand estimation approach to distributed PSS designers and decision makers. First, RP data provides overall results superior to decision making heuristics, crowdsourcing, and other traditional demand estimation methods. This advantage, however, must be balanced with the higher accuracy of traditional methods at the highest demand locations. The best approach for distributed PSS

operators is to identify the highest priority areas utilizing a low-cost, non-deterministic method such as local knowledge, followed by algorithmic assessment of the remaining locations.

Limitations of this approach include the fact that the decisions to ride may be conditionally dependent, nullifying one of the framework assumptions that make the binomial distribution applicable. For example, individuals who observe other people riding might be inspired to also choose to ride. Additionally, this approach assumes that the user population characteristics are constant over time. While applicable for this case study due to predicting utilization one year later, shifting user demographics may limit this approach's applicability to applications such as long-term infrastructure planning with decades look ahead time. Including these effects could further improve the power of the approach introduced in this thesis and allow for longer-term planning and influence of the initial design.

The approach to RP demand estimation of distributed PSSs should be tested in additional industries and design cases. Vending machine capacity should depend on the number of people considering buying a product (n) and the probability that they will choose buy something from the vending machine (p). Rubbish collection bin placement and capacity is also dependent on both the number of people with trash (n) and probability that they will not litter (p). Finally, proper implementation of this method may provide diagnostic insights into population behavioral changes unknown from average hourly utilization alone. For example, if a technician was unsure of the number of units in his/her service area, this process could help identify if increased service calls was due to an increased number of machines in his/her district (increased n) or a decrease in machine quality from the factory (increased p). This last type of application could prove to be a powerful Quality Assurance tool.

A key gap remains, however. What is the relationship between the calculated average *n* and average *p* surfaces to environmental features? Similar to the approach taken by Rixey, if the derived surface can be mapped to environmental features, then new surfaces can be accurately inferred from environmental features alone [57]. This inference will allow the surface to be accurately extended beyond the boundaries of the currently served area, increasing this approach's applicability. This gap will be investigated in Chapters Five and Six.

CHAPTER 5 Case Study 2: DIVVY Expansion Outside the Boundary: Background and Methods

Socio-technical Environmental Systems (STES) are defined by the interactions between *technical artifacts, user populations,* and their *environments*. The results of Case Study One enhanced understanding of the behavior of *a new user population* given a constant *technical artifact* over changing *environmental conditions*. Estimating binomial distribution parameters *n* (user population size) and *p* (user population product affinity) from historical user data allowed demand prediction in new situations. This approach was applied to a major Bike Sharing System (BSS) expansion. Plotting the estimated parameters revealed continuous Demand Surfaces over the BSS area, allowing prediction of overall ridership levels at new station locations. The results yielded a stronger correlation to the observed new station utilization (rho=.830,stations=46,p<.01) than the order implemented by the BSS operator (rho=.596,stations=46,p<.01), validating the approach of using current user data to estimate *user population* characteristics to informing design decisions in new *environments*.

Product service systems (PSS) are uniquely dependent upon timely or expensive user data for system planning, yet user datasets are only accurate for a small part of the entire PSS. Thus, methods to use the available data effectively and use data collected in one portion of a PSS for system design in another portion could transform PSS design. The designer faces a unique challenge when using system data to estimate PSS demand. PSS demand varies throughout the area serviced by the PSS and PSSs are often introduced in situations where user data is unavailable. Additionally, even when user data is available, demand estimation approaches were designed for traditional products, not PSSs.

Stated Intention (SI), Stated Preference (SP), and Revealed Preference (RP) demand estimation methods have significant challenges for product service systems. SP and SI approaches can require time-consuming or expensive surveys. SP and SI methods may be inadequate due to user's inability to accurately forecast their demand for a new service. SI results tend to overestimate demand due to self-selectivity bias, non-commitment, and exaggerations intentions to drive the overall results of the survey[9].

To avoid SI and SP, RP demand estimations may be created by analyzing comparable products. For completely novel products, designers may consider creating a test market to observe users; however this may not be practical for large PSS[7]. Without SI, SP, or RP, demand estimation methods such as local knowledge, expert guidance, heuristics, or "gut-feel of the decision maker" may be employed [7,10,11]. As a result, PSS design decisions may not be quantifiably repeatable or built from evidence[3]. For example, when 29 firms estimated the demand for wireless technology in the early 1980s, estimates varied by as much as 650%[7]. These challenges highlighted the need for a new RP approach for PSS demand estimation.

Case Study One presented an approach to predict the various level of user demand throughout an existing PSS service area. This work provided a PSS demand estimation starting point by successfully predicting PSS demand in new situations by estimating binomial distribution parameters n (user population size) and p (user population product affinity). These parameters provided a reliable prediction of future demand, but only in areas with available user data.

In this scenario, information is available about some users in a given context or scenario; however, the designer must derive an approach to allow him/her to apply the known user information in a new scenario. Insights into this type of problem could allow designers to better design new systems utilizing available limited data. For example, a framework to allow planners

in New York of a new telecommunications system to utilize user demand from other cities to frame their approach could limit demand estimation inaccuracies.

Although Case Study One demonstrated the ability to derive *n* and *p* fields from existing user data, this approach was limited due to only being applicable to the new stations within the current boundaries of the existing BSS. This type of *n* and *p* field will be referred to from this point forward as Localized Demand Surfaces (*n* and *p*). When the Localized Demand Surfaces were applied to the 128 stations existing outside the 2014 BSS boundaries, Spearman's rho dropped from 0.830 (n=46) to 0.33 (n=128) and the R^2 value for E[x] vs μ dropped from 0.207 to $8x10^{-5}$ (Figure 13). Although this was an improvement over the implemented operator ordering (rho=.146, n=128), extending Localized Demand Surfaces beyond the boundaries of the available user data significantly limited this approach's effectiveness.

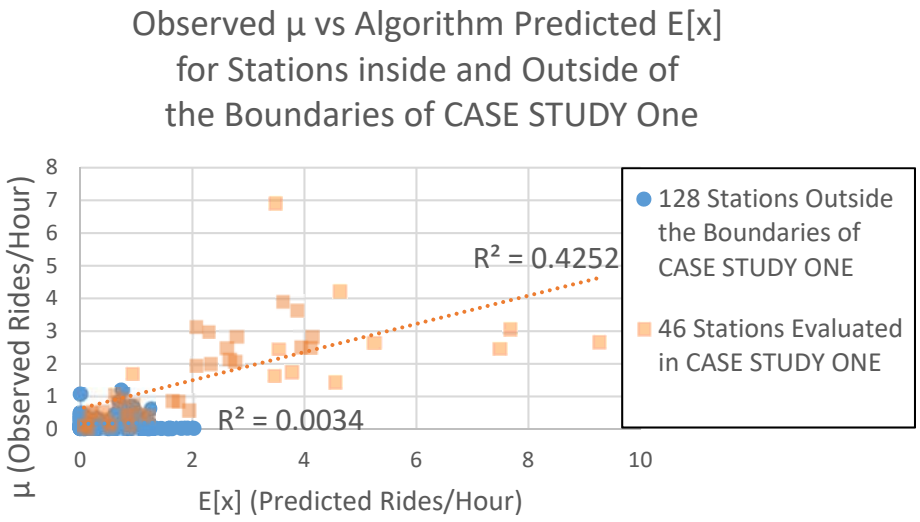


Figure 13: Reduction in Algorithm Accuracy Outside the Boundaries of Case Study One

The ineffectiveness of the model beyond the user data boundary may be due to *n* and *p* estimations in locations with limited ridership. Additionally, non-realistic boundary behaviors occur due to extending analysis beyond the boundaries of the data used to create the surface (Figure

14). Conversely, Spearman's rho (.32) was still higher than expected. The *n* and *p* surfaces generally tend to decrease as the distance from Lake Michigan increases, which may explain why the model developed in Case Study One has any predictive power in the area beyond the boundaries where it was built.

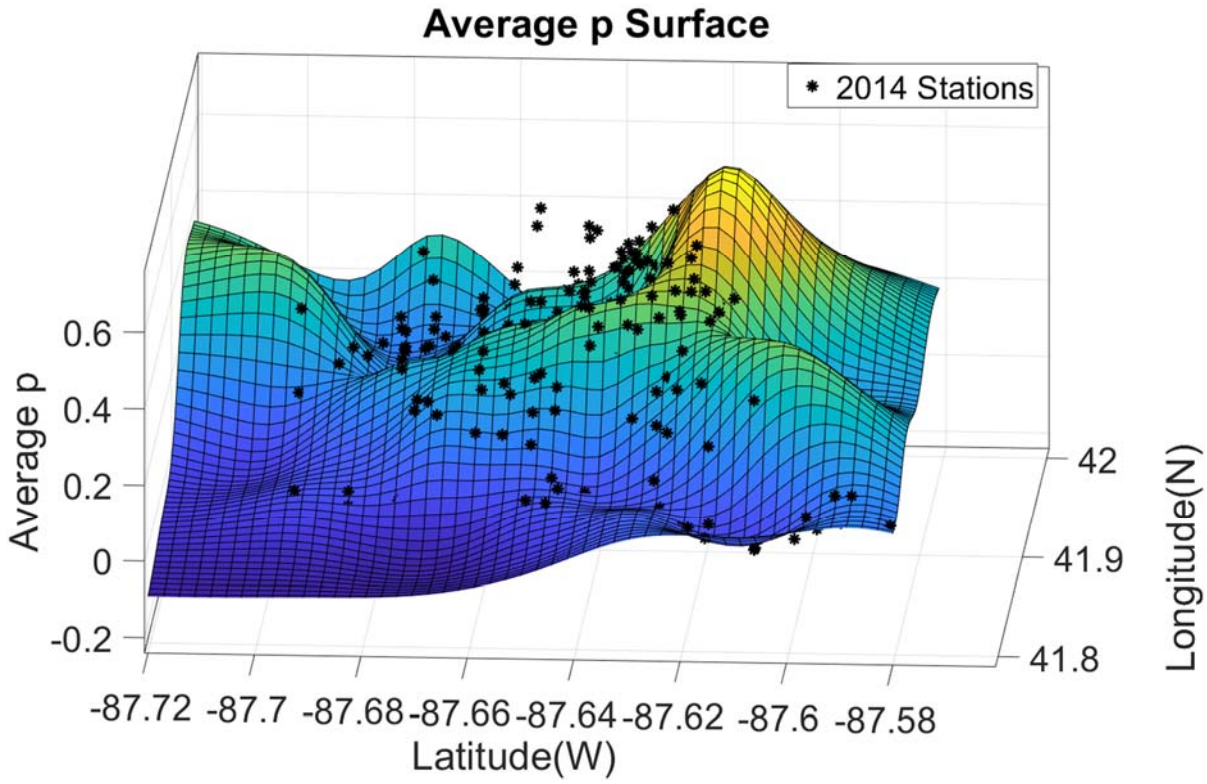


Figure 14: Example of Rapidly Increasing Boundary Conditions: Case Study One n Surface

The central question examined in Case Study Two is: *how can designers compensate for situations where the design environment has changed and limited user data is available to create SI, SP, or RP demand estimations?*

Given the following:

1) *n* and *p* are characteristics of the aggregate user population and environmental conditions.
2) User Data is unavailable to estimate user population and environmental characteristics in a new design context.
3) Other data sources exist which provide insight into aggregate user population and environmental characteristics.

Our hypothesis is that publicly available user population environmental characteristics data can be used to estimate Demand Surfaces outside of the boundaries previously constrained by available user data. These will be referred to as Regressed Global Demand Surfaces. This approach uses available Localized Demand Surfaces to determine the relationship between user population characteristics, environmental characteristics, and the magnitude of *n* or *p*. Once these relationships are determined, city-wide population and environmental variables can be used to create Regressed Global Demand Surfaces in areas without user data. Thus, this approach will fulfill a key gap of Case Study One's approach, providing designers insight into optimal station sizing in locations without available user data.

Case Study Two focuses on determining the relationship between the calculated *Demand Surfaces* to the *environmental and population characteristics*. This allows *n* and *p* for new locations to be inferred from environmental and socio-demographic variables alone, providing a method to link the *user population behaviors* and the *environmental conditions* when creating a STES model. The main tool used to investigate this question is multivariable regression. User Socio-demographic variables include factors such as income, race, and age. Environmental

variables include factors such as restaurants, terrain, and infrastructure conditions[3,11,56,57,61–64].

This approach uses two types of demand surfaces. First, Localized Demand Surfaces (*n* and *p*) are estimated (Case Study One). These surfaces exist *within the boundaries* of available user data. Next, regressions are used to discover the relationship between socio-demographic variables, environmental variables and the magnitude of *n* or *p* (the Localized Demand Surfaces). Once these relationships are determined, the designer can estimate the values *n* and *p* from environmental and socio-demographic variables alone. These new estimates create the Regressed Global Demand Surfaces.

This approach was validated by applying these surfaces to a major Bike Share System Expansion, outperforming the currently employed methods utilized by the BSS operators. The case study successfully demonstrates an approach for PSS design when designers do not have available user data for new locations.

The contributions of Case Study Two are as follows:

5. These results provide a framework to transform geographically limited available PSS user data into design insights for the portion of the system without user data.
6. This work provides a second validation of the value of *n* and *p* estimations for PSS planning as proposed in Case Study One.
7. This work provides a tool for BSS operators planning a system expansion.
8. This work identify environmental and socio-demographic variables that correlate with higher BSS use.

The remainder of this chapter is organized as follows. A background provides historical insight into the problem estimating demand in situations without user data and previous BSS

regression attempts. Methodology presents the data sources utilized for the regression, a brief description of the regression approach taken, and the four tests used to determine the applicability of the Regressed Global Demand Surfaces.

5.1 Background

5.1.1 The Problem of User Demand Estimation in New Situations: Insights from Disruptive Innovations

Incorporating anticipated user demand into PSS design is problematic when user data is not available for a new market or situation. Although one could collect a small amount of data and re-perform the analysis of Case Study One, this results in opportunity cost and may not be effective [66]. Much previous work on incorporating user demand when data is unavailable is within the field of Human-Computer Interaction or software development. This field easily allows for near final prototype testing and rapid design changes of the final product [67], not always possible for mechanical systems. Current applications include active utilization of user context for information delivery and game usability testing [67,68]. Methods such as user interviews, role playing, or user interactions with prototypes are used to guide the design process [5].

More insights into the investigation of design situations with limited user data question can be found in the current research on the adoption and spread of disruptive innovations. Disruptive innovations introduce a new technological advancement, dramatically affecting market demand by replacing current technologies [66,69]. Examples include Micro-Electrical-Mechanical devices (MEMS), Electric-bikes, the personal computer, cloud computing, advanced 3-D printers, and cellular telephones [66,69–72].

A variety of approaches have been proposed to predict demand of disruptive innovations. At the most basic, expert opinions are used. Linton refined this approach by using expert opinions

about projected supply and demand to inform Monte-Carlo simulations, allowing inclusion of expert uncertainty into his models [70]. Application of this approach, yields a probability distribution for expected demand [71]. Diffusion models have also been used to predict demand, however without data to estimate model parameters for the market being examined, accuracy suffers [66]. Useful user data is often only available after the disruption process has begun [69].

Identifying disruptive innovations before they shift the market place could provide dramatic benefits to companies [70]. The potential benefit is so large that authors note that even though additional algorithmic work is required, any advance in this field could be leveraged for significant profit [66,69–71]. As a result, many efforts have been made to simply pre-identify disruptive innovations, rather than predict the demand magnitude. This includes analysis of patent data, hazard function, text mining, and web mining [72–74].

Previous disruptive innovation research provides insight to the investigation of demand estimation beyond the boundaries of available user data. Disruptive Innovation demand estimation research provides a precedent for using past behavior to predict future demand [70]. Disruptive Innovation profitability also requires consideration of each market, analogous to socio-demographic and environmental variables in this thesis [66,70]. This provides a basis for the approach of using socio-demographic and environmental variables to directly estimate PSS demand in new situations. Researchers caution, however, against applying a model from one marketplace to another although evidence indicates it may be possible to adjust for differences within environments [66]. Thus, this study utilizes regression techniques and applies these estimations within the same marketplace (different parts of the same city). Conducive environmental variables such as bike paths or cultural considerations can be shown to increase the

adoption of disruptive technologies [75], providing justification for including environmental variables into this thesis.

Although similar, the previous work on disruptive innovations differs from the focus of this thesis in several key areas. First, this thesis examines the situation where some market data is available, while disruptive innovation research attempts to identify if an innovation is disruptive and then predict demand with no market data. User data may not be available because even though the technology is identified as disruptive, it may not be developed enough for market entry. When applying SI, SP, and RP demand estimation, this uncertainty can yield wide variance in demand estimation. For example, expert estimates for the MEMs market size in 2000 ranged from 2 billion to almost 30 billion dollars annually [71]. This is a fundamental difference between identification of disruptive products (disruptive innovation research) and successful implementation of products with limited user data (the second major question examined in this thesis). Secondly, disruptive innovation demand prediction often focuses on dynamic demand growth [70], while this research only examines final steady state demand. Third, disruptive innovation research often examines factors beyond the scope of this investigation, such as the role of friendly government legislature to the innovation [75].

Finally, examining an Innovative PSS may provide insights or overcome traditional challenges faced by disruptive innovation research. The challenge of accurately forecasting multiple sales to a single individual is not applicable to PSS [66]. For a PSS, designers are concerned about infrastructure use, and the usage per individual is not a vital statistic as for traditional products. Of note, this gap in repurchasing demand estimation may provide insight into why traditional SI, SP, and RP methods struggle to predict PSS demand.

5.1.3 CASE STUDY: Station Clustering Algorithms work in the literature.

Various methods have been historically employed to estimate areas of high demand when planning an initial BSS expansion or installation [3]. There are two distinct approaches to planning BSS station locations. First, experts generated a list of desirable environmental characteristics, analyzed their density within their target cities, and used them to create a BSS potential heat-map. A discussion of BSS planning in Philadelphia and Boise illustrates this first approach. These studies were used by city planners to determine new station location and sizing. Philadelphia planners reviewed the lessons from Montreal, Lyon, and Paris when identifying the nine variables to consider in 2010 [62]. Environmental features were given a weighting and were summed within 500 meters of potential station locations. Philadelphia planners, however, did not recommend station sizing. The authors recommended combining this approach with surveys and other qualitative methods[62].

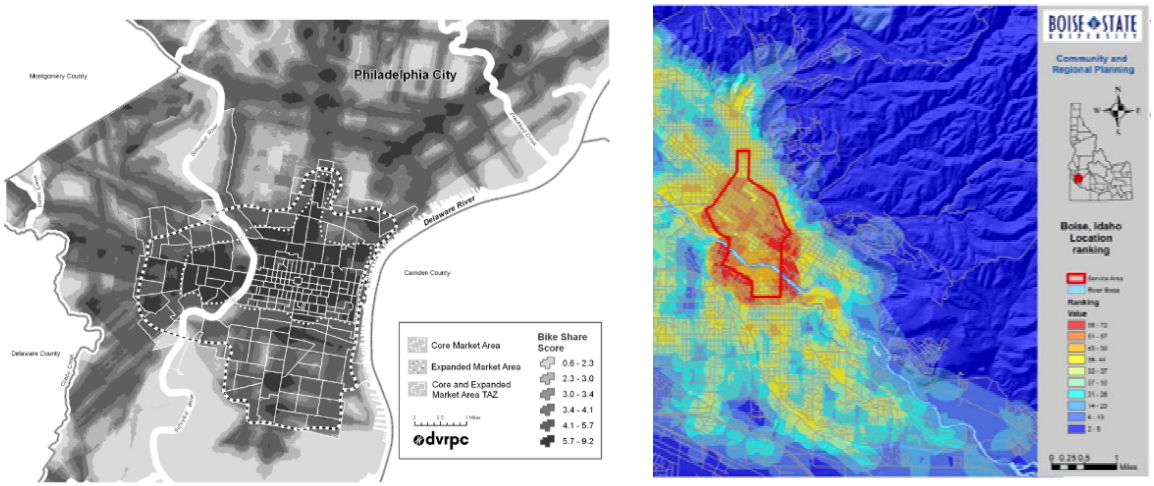


Figure 15: Example Early BSS "Heat" Maps generated by local planners [61,62].

Increasing in refinement from the approach taken in Philadelphia, Boise planners in 2012 considered 11 variables and two stations sizes when creating their BSS heat-map [61]. Of note,

due to Boise being generally flat, the researchers did not include topography considerations. This study incorporated bicycle accidents, reasoning that a high number of accidents implies a large volume of traffic. Thus, unexpectedly, bicycle accidents are positively correlated with ridership. The limitation of these two studies is that without available ridership data, no analysis was performed to assess the accuracy of the predictions for Philadelphia or Boise.

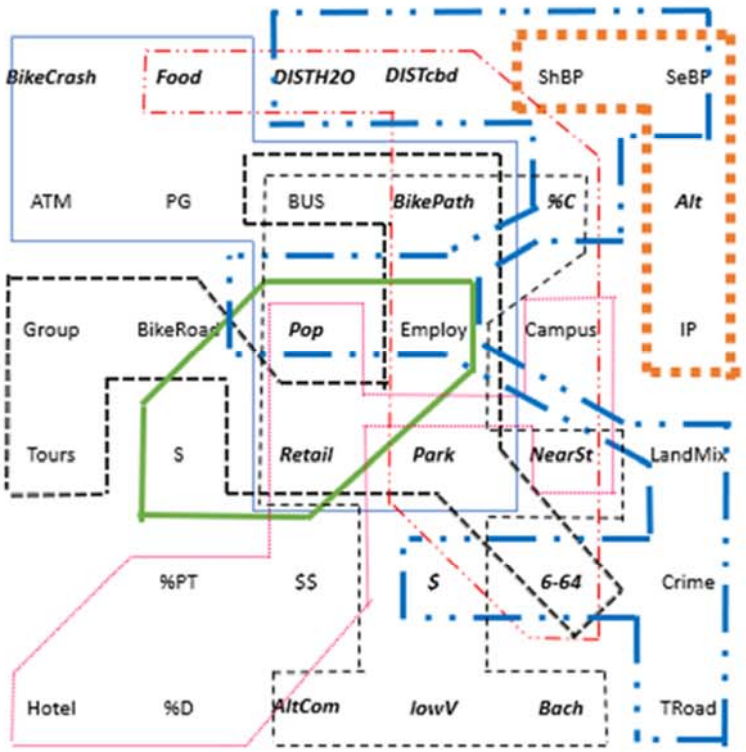
Once ridership data became available, a second generation of BSS environmental studies utilized observed ridership data to validate regressions conducted with environmental characteristics. As shown in Table 3 and Figure 16, these studies resulted in a wide variety of accuracies, independent variables, and approaches. A Brisbane study examined the role that infrastructure and environmental features such as elevation played in station usage. The resulting linear regression found an R^2 of 0.43 when considering these factors alone, highlighting the importance of a conducive cycling environment to BSS demand [64]. An analysis of pedestrian and bicycle traffic in Minneapolis incorporated additional socio-economic factors [3]. This study utilized through-traffic counts of all bicycle traffic and did not focus on BSS usage. Insights include the importance of appropriately varying the radius of influence for environmental variables when considering different transportation modes [3]. A related study analyzed car-share station placements in Nice, France incorporated additional socio-economic factors [11]. Finally, a correlation analysis at the end of an autonomous usage profile study provided evidence that BSS usage is correlated with population, jobs, services, and shops [56].

The previous studies demonstrated moderate correlations between subsets of environmental variables, socio-dempographic variables, and BSS demand. Studies that combined both environmental and socio-demographic variables resulted in very strong correlations. This approach was demonstrated in both Minneapolis St. Paul, Washington D.C., and Denver with R^2

greater than 0.8 [57,63]. These studies however, utilized monthly rentals or total demand, rather than *n* and *p* as done in this study. The models developed in Case Study Two uses significantly fewer variables than these two studies but with comparable accuracy.

Table 3: Pertinent Characteristics of Previous Environmental Regressions

City	Dependent Variable	Environmental Radius	Type of Model	R^2
Boise [67]	Points	820ft and 1640ft	Points for each possible location based on amenities within two distances.	N/A
Philadelphia [66]	Points	500M	Points for each possible location based on amenities	N/A
Brisbane [68]	Frequency of Station Usage	400M	Linear Regression	0.43
Paris [50]	None	Per Hectare	Performance profiles are related to environmental variables, no prediction or R^2 provided.	N/A
Minneapolis and St Paul [69]	Total Station Activity (arrivals and Departures)	400M (200M for food)	Log Linear OLS and Negative Binomial Regression.	.847 and .863
Nice[11]	Car Station Performance	500M	Linear Regression	"fairly robust and have reasonable measures of fitness"
Minneapolis and St Paul [69]	12hr rider counts (not BSS)	Census Block Group	OLS and Negative Binomial Regression	.381 and .476
DC, Denver, Minneapolis and St Paul [51]	Natural Log of Monthly Rentals	400M	Multivariate Linear Regressions	.802, .754, and .801



Mean Ridership	μ	Distance from Water	*DISTH2O*	Group Quarters	*Group*
Number of Riders	*n*	Number of Stations within 4800 ft	*St<4800*	Bike Friendly Road	*BikeRoad*
Probability of Riding	*p*	Distance to Nearest Station	*NearSt*	Employment	*Employ*
Average Number of Riders	*AVGn*	Total Population	*Pop*	Campus	*Campus*
Average Ride Probability	*AVGp*	Median Income	*$*	Informal Path	*IP*
Bike Path Length	*BikePath*	Percent Caucasian	*%C*	Tourist Attraction	*Tours*
Shared Bike Path	*ShBP*	Population 6-64	*6-64*	Services	*S*
Separate Bike Path	*SeBP*	Number of Low Vehicle Individuals	*lowV*	Land Use Mixture	*LandMix*
Number of Restaurants	*Food*	Number of Bachelor's Degrees	*Bach*	Percent Public Transit	*%PT*
Number of Retail Locations	*Retail*	Number of Alternate Commuters	*AltCom*	High Income	*$$*
Number of Bike and Pedestrian Crashes	*BikeCrash*	ATM	*ATM*	Crime	*Crime*
Number of Park Attractions	*Park*	Parking Garages	*PG*	Hotel	*Hotel*
Number of Train Stops	*Train*	Bus Stops	*BUS*	% Drivers	*%D*
Distance from Center of Business District	*DISTcbd*	Elevation	*Alt*	Type of Road	*TRoad*

Figure 16: Previous Studies that Predicted Ridership with Environmental features incorporated a wide variety of independent variables
Note: The most frequently used independent variables are in the center of the grid.

5.2 Methodology

The following is the overall approach for utilizing binomial parameter estimation with environmental regressions for estimating PSS demand outside the boundary of existing user data. Major steps and considerations are as follows.

1. Gather Demand Data: User data is required for the initial estimation of the Localized Demand Surfaces. Considerations include appropriate data discretization and minimizing the possible values of *n* to increase estimation accuracy.

2. Create Localized Demand Surfaces (Figure 10). These surfaces provide designer insight for increasing the density of PSS. Additional details and examples of steps 1 and 2 are provided in Chapters 3 and 4**.**

3. Gather Environmental and socio-demographic variables: Once Localized Demand Surfaces are generated, the next step is to gather potentially relevant variables including both socio-demographic (income, race, gender) and physical environment (infrastructure, climate, terrain). These data sources could be selected after review of current demand estimation approaches and consultation of system experts.

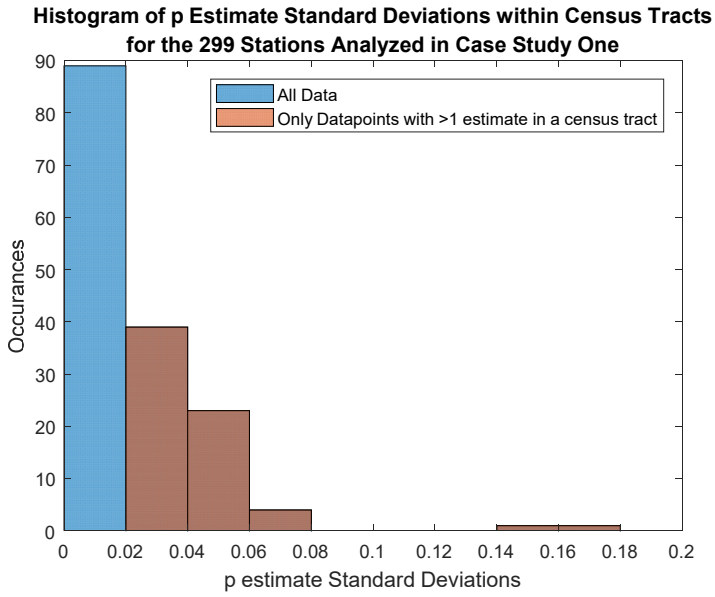
4. Calculate Multiple Linear Regressions: To create Regressed Global Demand Surfaces, multiple linear regressions are calculated to relate the Localized Demand Surfaces to the environmental and socio-demographic variables gathered in step 3. The purpose is twofold. First, it enables determination of which variables are relevant. Secondly, the final regressions are used to create the Regressed Global Demand Surfaces. When accomplishing this step, this thesis utilized a step up approach with multi-collinearity checks.

5. Generate Regressed Global User Characteristic Surfaces to predict demand at new locations: Once the regressions are determined in step 4, the environmental and socio-demographic variables outside the Localized Demand Surfaces is used to calculate n and p for new locations throughout the PSS. n and p then allow calculation of the expected hourly utilization, E[x].

$$E[x] = n * p \tag{13}$$

Next, steps 3 and 4 are illustrated through the analysis of the DIVVY BSS expansion case study.

5.2.1 Step 3: Gather Environmental Data to Create Regression Dataset



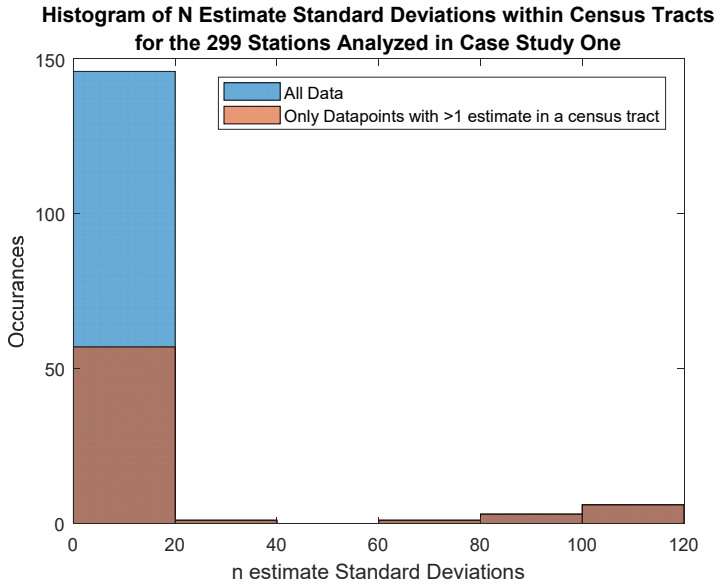


Figure 17: Wide Variation of n and p within a single census track.
Note: All measurements above the smallest standard deviation are identical for all data and data points with >1 estimate in a census track

An initial investigation of *n* and *p* from Localized Demand Surfaces revealed wide fluctuations of *n* and *p* within the same census tract (Figure 17). Greater than one estimate refers to locations with more than one *n* estimate available to create the weighted *n* average value assigned to that census track. Clearly, census tracks with only one *n* or *p* estimate would have a standard deviation of zero. This resulted in the large peaks in the smallest histogram bins.

Due to the wide variation within census tracks, two types of input variables were desired. First, variables that continuously varied through the census tract and were not derived from census data. This includes environmental characteristics such as bike path length, retail, and restaurants. This first type of data still excludes valuable socio-demographic variables collected via census. Thus, a second type of approach was required to enable utilization of census data. This approach created data that was the weighted average (by relative area) of all census tracks within a defined

radius of the potential BSS station locations. This includes socio-demographic variables such as race and income.

Table 4 outlines ten independent variables that were gathered for analysis based on previous demand estimations in the literature. The mean values are for the 300 stations analyzed in Case Study One to create the Localized Demand Surfaces. Two influence radiuses were utilized. 2.57 km (1.6 miles) was used for crashes, parks, and paths. 2.57 km (1.6 miles) is the average ride length and these features were more likely to be enjoyed or experienced while riding [41]. 0.4 km (0.25 mile) was used for all destination attractions to be consistent with last-mile literature [59]. Park attractions were used instead of square footage or number of parks because it should more closely correlate with usage. A small park with many attractions may generate more BSS demand than a large, empty park. Attractions include features such as playgrounds, basketball courts, and restrooms. Bus stops were considered, but not analyzed due to the high density within the 2014 BSS station area. All stations would have had approximately the same number of bus stops within their influence radius. The center of the business district and line approximating the shore of Lake Michigan were estimated by the researcher. The central business district was defined by Chicago City Scape (https://www.chicagocityscape.com/maps/index.php?place=custom-central-business-district), a website used by developers and the construction industry to track construction and permit information for the city of Chicago.

Table 4: Environmental Independent Variables

Variable	Units	Mean and Range	Radius (miles)	Source	Description	Literature Precedent
Bike Path Length	Miles	.0192 (.024-.160)	1.6	Chicago Data Portal[76]	Total Length of Paths within Radius	[3,57,61–64]
Number of Restaurants	N/A	34.77 (0-169)	.25	Chicago Data Portal 2014 Food Inspections	Total number of Restaurants within radius	[61,63]
Number of Retail Locations	N/A	26.42 (0-80)	.25	Chicago Data Portal 2014 Annual Taxpayer Location List	Total number of Stores within radius	[11,56,61,62]
Number of Bike and Pedestrian Crashes	N/A	21.33 (0-55)	1.6	Chicago Data Portal	Number of Pedestrian and Bikes hit by cars.	[61]
Number of Park Attractions	N/A	170.61 (85-298)	1.6	Chicago Data Portal	Total Number of Attractions	[57,61–63]
Number of Train Stops	N/A	1.51 (0-16)	.25	Chicago Data Portal	Total number of Train stops.	
Distance from Center of Business District	Miles	2.76 (.12-7.15)	N/A	Calculated	Center of Business District Defined as 41.89N/87.63W.	[3,63]
Distance from Lake Michigan	Miles	2.30 (.103-6.42)	N/A	Calculated	Lake Michigan defined as line along 41.999N/-87.654W and 41.765N/-87.559W	[3,63]
Number of Stations within 4800 ft	N/A	23.37 (2-54)	.91	Measured	Number of Stations within 4800 ft	[57]
Distance to Nearest Station	Miles	.2377 (.08-.84)	N/A	Measured	Distance to Nearest Station	[11,63]

The second type of data collected for analysis was aggregated Census Data. Due to many of the BSS stations existing on the edge of a census track, the census values were weighted and averaged within a 0.4 km (0.25 mile) radius. The raw data was received from www.AmericanFactFinder.com's 2014 American Community Survey with 5 year look ahead (reports B01003, B02001, S1903, B08201, S0601, S0801). The data analysis was performed in ArcGIS 10.3.1. This resulted in an additional seven independent variables for analysis. Number of Caucasians was also considered instead of percent Caucasian, but no significant correlation was discovered.

Table 5: Socio-Demographic Independent Variables Evaluated

Variable	Mean (Range)	Literature Precedent
Total Population	3,923 (808.5-12,468)	[3,11,56,57,61,64]
Median Income	75,355 (18,535-138,250)	[3,57]
Percent Caucasian	66.78 (.46-93.88)	[3,57,63]
Population 6-64	286 (12.43-1,587)	[3,62,63]
Number of Low Vehicle Individuals	700 (23-4981)	[57]
Number of Bachelor's Degrees	797.5 (24-3,094)	[57]
Number of Alternate Commuters.	1,295 (116.1-5,214)	[57]

Tables 4 and 5 provide a combined 17 independent variables for analysis. The dependent variables were the average *n* and *p* surfaces generated by the 299 stations existing in 2014. μ was also included as a dependent variable to allow for comparison with previous studies that focused on estimating hourly ridership, not *n* or *p*. Additionally, this allowed direct testing if regressions to estimate *n* and *p* provide additional information and accuracy from a regression for μ alone.

5.2.2 Step 4: Calculate Multiple Linear Regression

Once the independent and dependent variable are collected into a single dataset, they are analyzed to determine appropriate linear regressions. Pairwise linear correlation coefficients and associated p-values are presented in Appendix 1. The variables in Tables 4 and 5 were normalized by dividing each datapoint by the mean of their dataset. To identify appropriate dependent-independent variable matching, 0.4 correlation was chosen for μ and p, while 0.3 was chosen for n. The *n* cutoff was lower, due to the weaker general correlation all independent variable showed with n. Variable were examined for multi-collinearity by utilizing the variance inflation factor (VIF), with VIF greater than ten being eliminated from consideration [77]. Next, a step-up approach was used to create the model and ensure all relevant independent variables were included. Finally, a second multi-collinearity check was performed to ensure the final model was accurate.

5.2.3 Procedure to Validate results

Once the regression is completed, Regressed Global Demand Surfaces are generated. Four tests were designed to validate this approach, providing different scenarios to test the efficacy of the Regressed Global Demand Surfaces to predicting PSS demand. As in Case Study One, Testing is defined as using the comparing both the algorithm and implemented operator ranking it to the ideal ranking with spearman's rho. There are two metrics used to evaluate the success of the algorithm's predictions. First, as in Case Study One, Spearman's rho can be calculated for both the implemented ranking and algorithm suggested ranking. Secondly, Pearson's correlation can be calculated between the observed hourly ridership and the predicted hourly ridership or selected station capacity. A successful test is one where the Algorithmic approach results in a higher Spearman's rho and Pearson's correlation than the traditional approach taken by the operators.

5.2.3.1 Scenario 1: Testing the 46 Stations Within the Boundaries of the Available User Data (Case Study 1).

This area (shown with circle-1s on Figure 18) is expected to have higher ridership than those outside the boundary of the Localized Demand Surfaces. Thus, accuracy in this area may dominate overall system performance. Additionally, this test is contained within the boundary of the Localized Demand Surfaces used to perform the regression that is the basis for the Regressed Global Demand Surfaces. The predictions from the Regressed Global Demand Surfaces are compared to the Localized Demand Surfaces (the results of Case Study One) and the implemented operator ordering.

5.2.3.2 Scenario 2: Testing the 128 Stations Outside the Boundaries of the Available User Data

This section, circle-2s on Figure 18, is expected to have lower ridership per station than those within the boundaries of available user data. Additionally, this test is not contained within the boundary of the Localized Demand Surfaces used to perform the regression that is the basis for the Regressed Global Demand Surfaces. This is the key test of the hypothesis that publicly available socio-demographic and environmental variables can be used to estimate Demand Surfaces outside of the boundaries constrained by user data.

5.2.3.3 Scenario 3: Testing all 174 Stations with the Regressed Global Demand Surfaces

All stations added in the 2015 Divvy BSS Expansion, circle-1s and circle-2s on Figure 18, are tested with the Regressed Global Demand Surfaces to assess the

overall implementation in areas of both high and low ridership. The accuracy of the Regressed Global Demand Surfaces is compared to the implemented operator ordering.

5.2.3.4 Scenario 4: Testing all 174 Stations with both types of Demand Surfaces

Finally, all stations added in the 2015 Divvy BSS Expansion are tested with the curve best suited to that area. This should maximize algorithm accuracy. The Localized Demand Surfaces are used to predict ridership within the boundaries of available user data (circle-1 on Figure 18), while the Regressed Global Demand Surfaces are used to predict the ridership at the stations marked circle-2 on Figure 18. This test combines the results of Case Study One and this Case Study as a potential best practice for BSS operators considering an expansion

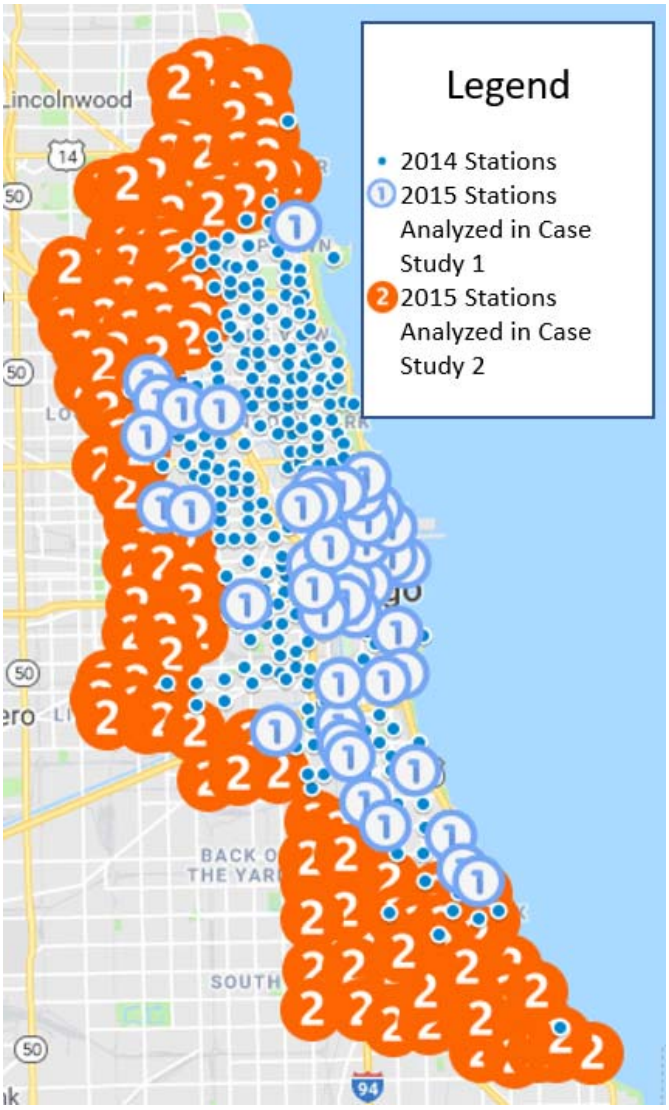


Figure 18: 2015 Station Expansion Subsets considered in the four tests in Case Study Two

5.3 Summary

Chapter 5 motivated the need for a second approach to applying *n* and *p* estimations for Product Service System Expansion. A background provides historical insight into the problem estimating demand in situations without user data and previous BSS regression attempts. Methodology discussed the data sources utilized for the regression, a brief description of the

regression approach taken, and the four tests used to determine the applicability of the Regressed Global Demand Surfaces. Next, Chapter 6 presents the results and limitations of these four tests.

CHAPTER 6: Case Study Two DIVVY Expansion Outside the Boundary: Results

Chapter 6 presents the results from Case Study Two. First, the regression is presented and analyzed. Then, the results of the four tests from Chapter 5 are presented, followed by concluding statements on limitations and necessary future work.

6.1 Regression Results

Table 6 summarizes the independent variables that exceeded the correlation and multi-collinearity thresholds for *n, p,* and μ. 0.4 correlation was chosen for μ and *p,* while 0.3 was chosen for *n.*

Table 6: Resulting n ,p, and μ Regressions

Variable	Coefficient			Standard Error			P Value		
	n	*p*	μ	*n*	*p*	μ	*n*	*p*	μ
Intercept	.480	2.43	2.89	.0757	~	.227	<.01	<.01	<.01
Bike Paths	.262	~	~	.0694	~	~	<.01	~	~
St<4800	.096	-.335	~	.0554	.0586	~	<.1	<.01	~
Median Income	.162	~	~	.113	~	~	<.01	~	~
Dist H20	~	-.517	-.533	~	.0315	.077	~	<.01	~
BikeCrash	~	.405	~	~	.0404	~	~	<.01	~
Dist CBD	~	-.195	~	~	.0539	~	~	<.01	~
Park	~	-.845	-1.58	~	.0956	.201	~	<.01	<.01
Food	~	.0599	.223	~	.0218	.0431	~	<.01	<.01
R^2	.178	.827	.457						

Legend

Intercept – Independent Variable intercept when all dependent variables are zero.
Median Income – Population Median Income within 0.25 mile radius.
St<4800 - Number of Stations within 4800 ft
Dist H20- Distance to Lake Michigan
BikeCrash – Number of Pedestrians and Bikes hit by cars within 1.6 mile radius.
Dist CBD – Distance from Center of Business District.
Park – Number of Park Attractions within 1.6 mile radius.
Food – Total Number of Restaurants within 0.25 miles.
R^2 – Pearson's Correlation Coefficient.
~ - Not included in regression due to less than correlation threshold.

Each regression coefficient was examined to ensure the direction of correlation was as expected. Of note, current literature examines correlation with μ, rather than *n* and *p,* making direct comparison difficult. Summary shown in Table 7. Correlation was as expected, with the exception of the park amenities variable. There is a negative correlation coefficient of park amenities to *p* and μ (-.845 and -1.58). Previous regressions utilized distance to park instead of park amenities, resulting in 0.061, -,486, and -.485 [57,63]. The negative coefficient might be because although more individuals choose to ride through the parks, the existence of large parks reduces the population density in those areas, resulting in an overall negative relationship. Additionally, docked bikeshare may be used more for commuting, thus recreational amenities have relative lower appeal, resulting in an overall negative correlation.

The final regression for *p* yielded a strong correlation ($R^2 = 0.827$). The *n* regression failed to provide more than a weak correlation ($R^2 = 0.178$). Although there was evidence of correlation between the predicted *n* values and the evaluated independent variables, strong predictors were not discovered within the 17 evaluated variables. Number of Train Stops, Total Population, Population 6-64, number of low vehicle individuals, number of bachelor's degrees, and number of alternate commuters only showed an influence at high values forcing *n* to a stable value.

Table 7: Expected and Actual Regression Correlations

Variable	Calculated Correlation		Expected Correlation		Reason	Literature Precedent for Correlation with μ
	n	*p*	*n*	*p*		
Median Income	+		+		Expect a higher number of people to be interested in BSS usage as income rises.	-[3]
St<4800	+	-	+	-	The higher station density, the more number of people will be interest in using BSS, but also the probability that they choose to use that station will decrease.	
Dist H20	+	-	+	-	The closer to lake Michigan the higher the population density, but also the less likely people are to use BSS due to other obligations (work, shopping, etc).	+[3] -[63]
Crash		+		+	Crashes are indicative of an area with a high amount of bike traffic.	
Dist CBD		-		-	Same logic as distance to water	-[3] -[63]
Park		-		+	The more amenities the more likely people are to use BSS.	See Discussion Above
Food		+		+	Restaurants are an attraction for BSS riders.	+[63]

I attempted to include the influence of these variables by incorporating heuristics into the prediction algorithm, but overall model results did not appreciably change, thus they were omitted to minimize model complexity. It is possible that these variables only appear to drive *n* to a stable value as an artifact of the small size of the dataset evaluated. Two of these variables are shown in

the scatterplot in Figure 19. The diagonal shows a histogram for each variable, representing the distribution of values each variable had in the dataset.

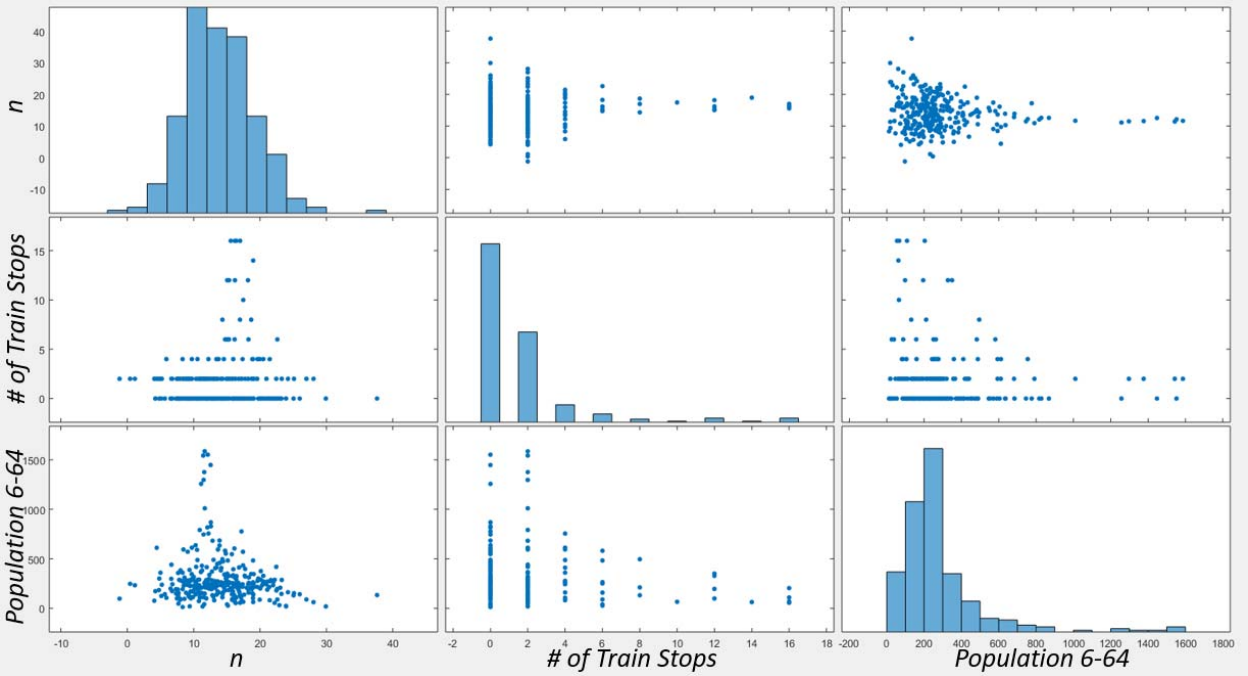


Figure 19: Scatterplot showing examples of two variables that asymptotically drove n to a stable value

The regression for μ resulted in only an R^2 of 0.457 rather than greater than 0.8 as seen in previous regressions [57,63]. Previous regressions, however, utilized monthly rentals or total demand, rather than average hourly demand as done in this study. Additionally, one study looked at total arrivals and departures while our approach just looks at outgoing demand. Also, these studies examined different cities, whose BSS utilization may be inherently more sensitive to environmental and socio-demographic variables. Finally, there is concern about the multi-collinearity of the models presented in previous works. For example, in the dataset examined in Case Study Two, high levels of correlation were observed between several of the variables used in these studies such as Alternate Commuters and Bachelor's Degrees (.9538) and Alternate Commuters and Population (.9185).

6.2 Test Results

Summary statistics and overall test results are presented in Table 8. A discussion of these results is presented in the following sections. There are two metrics used to evaluate the success of the algorithm's predictions. First, as in Case Study One, Spearman's rho can be calculated for both the implemented ranking and algorithm suggested ranking. Secondly, Pearson's correlation can be calculated between the observed hourly ridership and the predicted hourly ridership or selected station capacity.

Table 8: Summary of Results for Tests 1-4

Scenario	**One**				**Two**			**Three**		**Four**	
Predictor	LDS	RGDS	OP	μ	RGDS	OP	μ	RGDS	OP	COMBO	OP
Spearman's rho	.83	.74	.59	.75	.32	.14	.16	.56	.33	.55	.33
Pearson (R^2)	.43	.46	.38	.39	.04	.19	.01	.47	.38	.49	.38

Legend:

LDS- Localized Demand Surfaces
RGDS – Regressed Global Demand Surfaces
OP- Actual Operator Choices
μ - Surfaces built from μ regression (instead of *n* and *p*)
COMBO- Utilizing Both LDS and RGDS

6.2.1 Scenario One Results: The 46 Stations Evaluated in Case Study One

The Regressed Global Demand Surface was used to predict the ridership for the 46 stations existing within the boundaries of the 2014 existing stations. These are the same stations predicted in Case Study One. Algorithm performance slightly degraded from the Localized Demand Surface (rho = 0.83, stations=46, $p < .01$ to rho=0.74, stations=46, $p < .01$). As expected, observational user data creates superior performing Demand Surfaces than regressions. This is because the

regressions were built from the surfaces created by observational user data. Predictions from Regressions therefore included an additional source of error (regression inaccuracies) that were not included in the surfaces created directly from user data.

The μ regression built only from station average hourly ridership rather than *n* and *p* performed comparably with the Regressed Global Demand Surface (rho =.75 stations =46, p<.01). All three algorithmic approaches outperformed the implemented operator ordering's moderate correlation (rho=.59, stations=46, p<.01).

To provide a qualitative assessment of algorithm accuracy, instances where algorithm or operator ordering varied from ideal ordering by more than nine places (1/4 of the sample size) were assessed. This occurred fifteen times in the algorithm ranking and eighteen times for the operator ranking. Instances where predicted ordering was within four (1/10 of the sample size) of ideal ordering were also counted. This occurred sixteen times for the algorithm ranking and thirteen times for operator rankings. For a detailed analysis of the Localized Demand Surfaces, refer to Chapter 4.

As in Case Study One, the Implemented Operator Ordering is superior at identifying the highest ranked stations, indicating that combining this approach with current methods might be the best practice for decision makers. Figure 20 shows the error for each prediction from the operators and the algorithms vs the ideal ordering. A running average (every 5 predictions) is plotted to allow easier detection of overall trends. 1/4 and 1/10 sample-size error lines are overlaid to provide a sense of scale for the accuracy of the predictions.

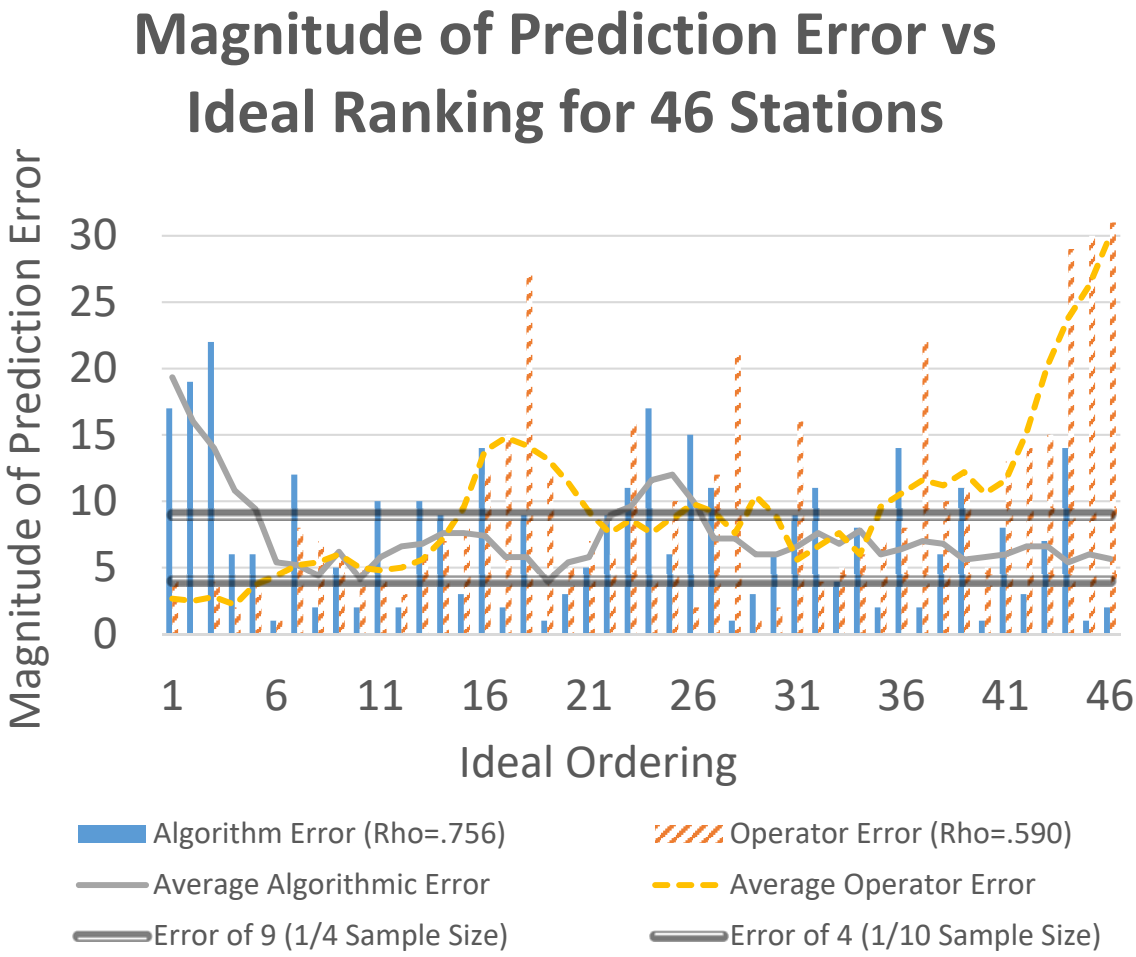


Figure 20: Running Average and Individual Prediction Error for the Environment Derived UCS and the Implemented Operator Ordering for 46 Stations

6.2.2 Scenario Two Results: the 128 Stations Outside the Boundaries of Case Study One

Next, the utilization of the remaining stations added in 2015 were predicted using the Regressed Global Demand Surfaces. The algorithm ordering showed a weak correlation to the ideal ordering (rho=0.32, stations=128, $p< .01$), but outperformed the implemented operator ordering, which was only very weakly correlated (rho=0.14, stations=128, $p<.1$). The μ regression

results degraded to rho=.16. The extra information encoded in determining both *n* and *p* instead of just μ doubled Spearman's rho.

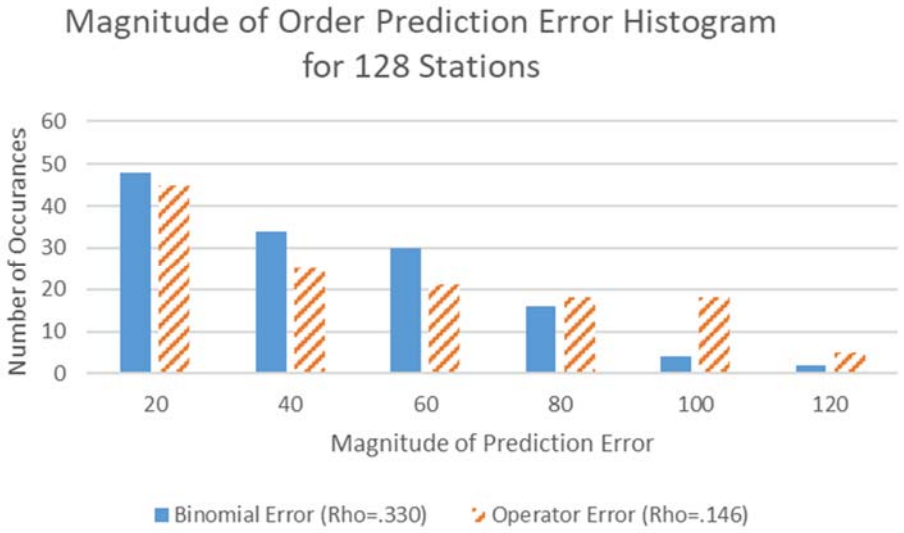


Figure 21: Histogram of Algorithm and Operator Error

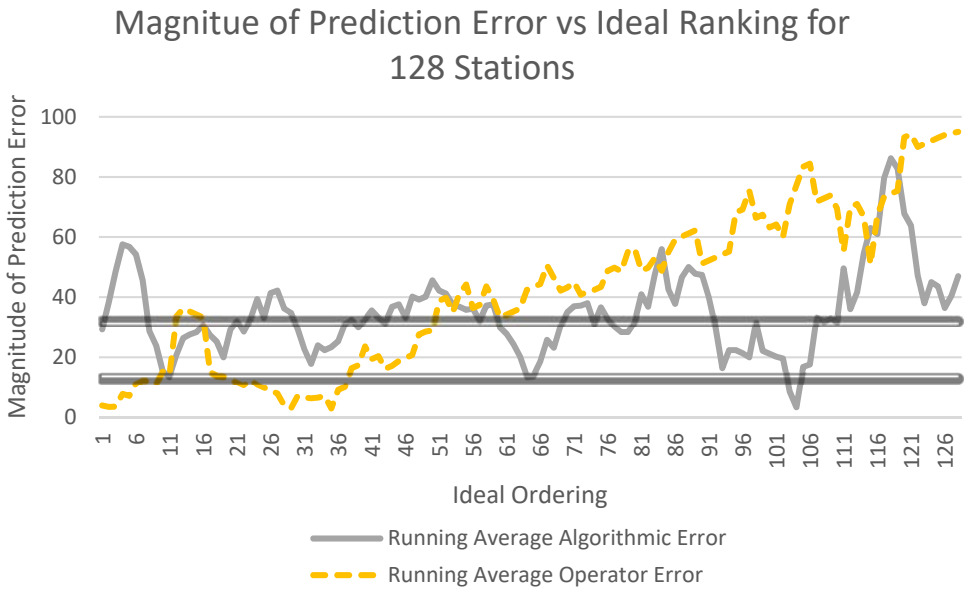


Figure 22: Algorithm and Operator Prediction Error vs Ideal Ranking

Figure 22 shows the trend previously observed in Case Study One still applies for this investigation: the implemented operator ordering methods are generally supperior in estimating

the demand of the highest performing stations, while the Algorithm performance does not appreciably degrade with ideal ordering. To maintain graph readability, only the running average is plotted. The individual predictions are ommitted, but this data is sumarized in Figure 21.

6.2.3 Scenario Three Results: all 174 Stations with only Regressed Global Demand Surfaces

Next, all 174 BSS stations added during the 2015 BSS expansion were tested and a station ordering was generated from the Regressed Global Demand Surfaces. The algorithm ordering showed a moderate correlation to the ideal ordering (rho=.56, stations=174, p< .01) while the implemented operator ordering was weakly correlated (rho=.33, stations=174, p<.01).

Figure 25 shows that the implemented operator ordering methods are generally supperior in estimating the demand of the highest performing stations, while the Algorithm performance does not appreciably degrade with ideal ordering. Figure 23 summarizes the accuracy of the algorithm and operator predictions.

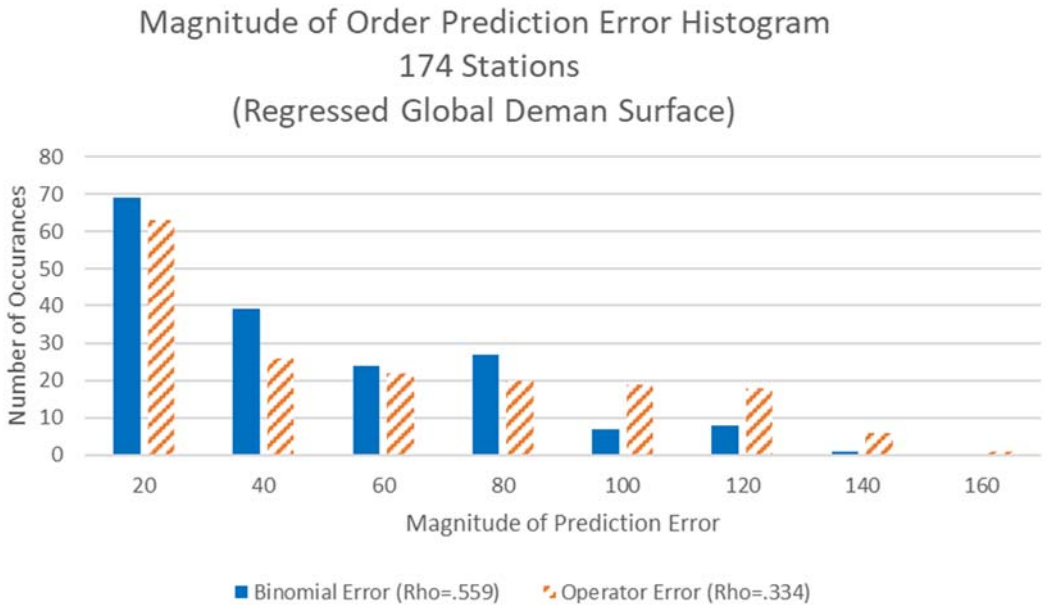


Figure 22: Histogram of Algorithm and Operator Error

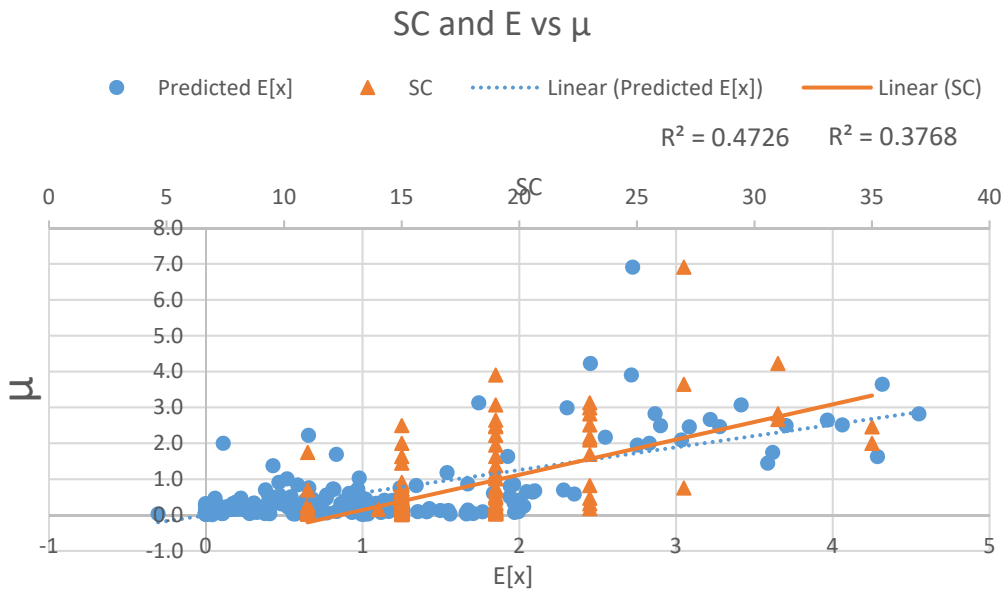


Figure 23: Regressed Global Predicted E[x] and Station Capacity vs Observed μ

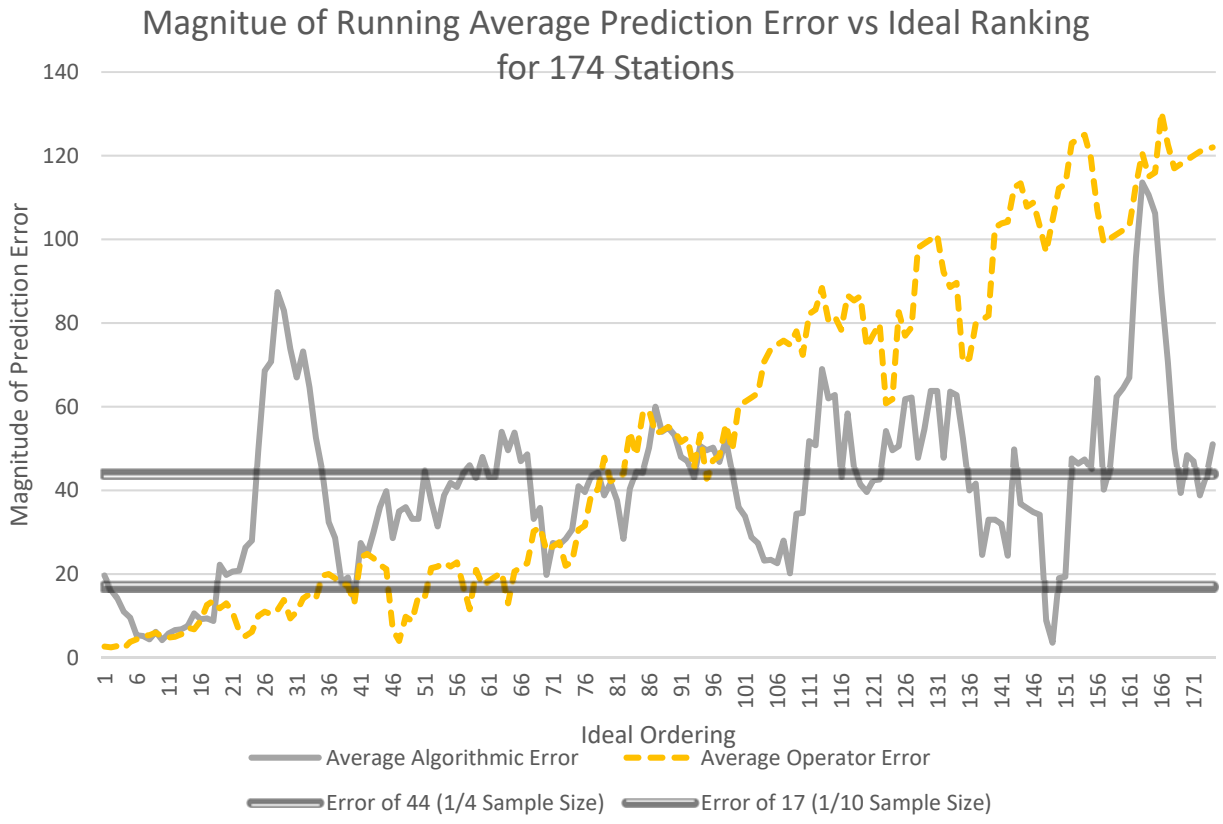


Figure 24: Running Average Algorithm and Operator Prediction Error

Both the operator implemented station capacity and algorithm expected average hourly ridership (E[x]) can provide operators insight into observed average hourly ridership (μ). As shown in Figure 24, both are weakly correlated, the expected average hourly ridership shows a stronger correlation with observed average hourly ridership (R^2=0.47, stations=174, p<.01) than selected station capacity (R^2=0.38, stations=174, p<.01). This indicates that the algorithm output is not only a better tool for ordering station sizes, but also selecting station sizes.

6.2.4 Scenario Four Results: all 174 Stations with both types of Demand Surfaces

The results of Case Study 1 and Tests 1-3 of this case study indicate that the optimal choice for BSS operators considering an expansion would be to use observational data to create Localized Demand Surfaces to estimate new demand for stations within the boundaries of the observational data (Case Study One). Then, Regressed Global Demand Surfaces should be created and used to estimate station demand outside the boundaries of the observational data.

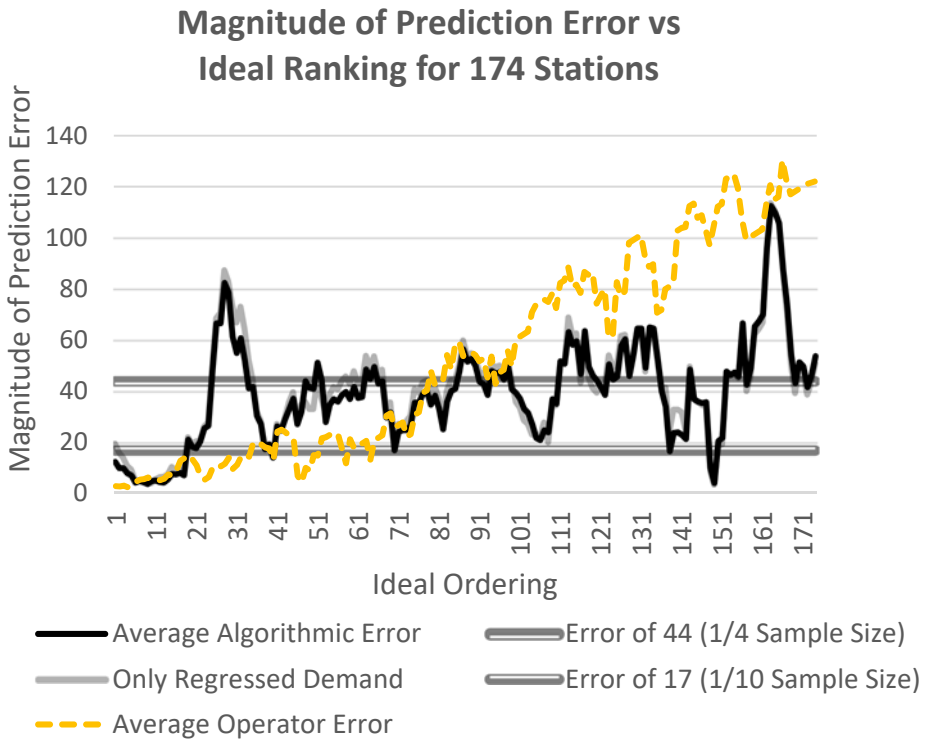


Figure 25: Running Average Algorithm and Operator Prediction Error vs Ideal Ranking for all 174 stations

When applied to this scenario, the resulting multi-approach algorithm ordering shows a moderate correlation with the ideal ordering (rho=.55, Stations =174, p< .01), while the implemented operator ordering was only weakly correlated (rho=.33, Stations =174, p< .01). Incorporating the Localized Demand Surface for the stations within the boundary of available user data did not appreciably change Spearman's rho. Excitedly, this indicates that if Regressed Global Demand Surfaces can be expanded further (from city to city vice simply different areas within a city), it might prove to be equally as powerful as Localized Demand Surfaces. This could allow BSS operators to omit costly or time-consuming data collection.

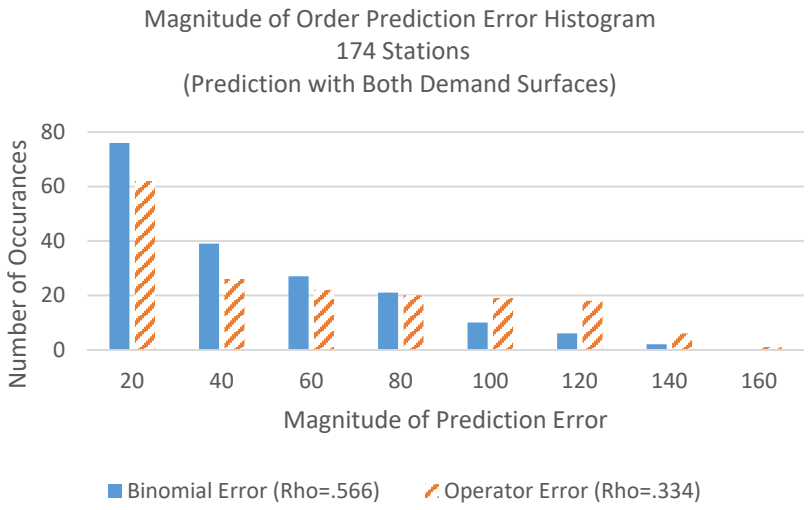


Figure 26: Histogram of Algorithm and Operator Error

When assessing the correlation between the predicted hourly station utilization and observed hourly utilization for the environmental conditions selected in Case Study One, the algorithm results in a significantly stronger correlation (R^2=.580, stations=174, p<.01) than the implemented station capacities (R^2=.377, stations=174, p<.01). As shown in Figure 28, this is also a E[x] R^2 improvement of nearly 0.05 over the results of Scenario 3.

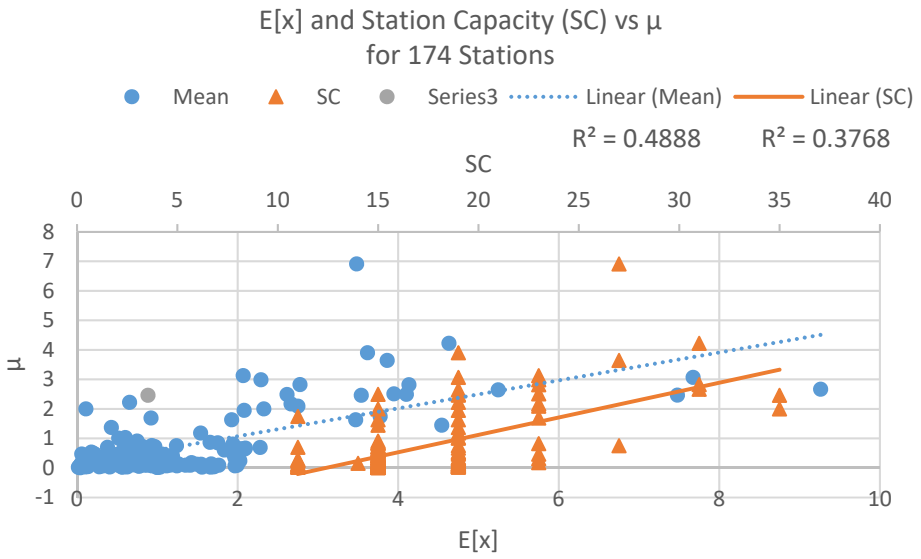


Figure 22:Combined Regressed and Localized E[x](rides/hour) and Station Capacity vs Observed μ (rides/hour)

6.3 Summary and Limitations

The analysis of four tests of this case study have demonstrated that the approach proposed in Case Study Two provides a partial solution to the research question: how do you transform available user data into demand estimations for new situations? Through the analysis of the 2015 Divvy BSS expansion, Regressed Global Demand Surfaces were used to better estimate demand for 128 stations existing outside the boundaries of the Localized Demand Surfaces than traditional demand estimation methods. This approach improved total BSS expansion algorithm Spearman's rho from 0.333 to 0.49 when using the Regressed Global Demand Surfaces.

Research gaps remain, however. The efforts in this thesis have examined demand estimation in a market with enough user data available to estimate Localized Demand Surfaces. In the BSS case study, the current results can only be applied to a BSS expansion. What about scenarios where only user data from a completely different environment is available? Can the results of this study be used to help BSS operators plan a system installation in a city that currently

has no BSS? Does this approach provide a tool for design across markets? These questions will be examined in future work. Additional future work could involve refining a method to combine operator ranking with the results of the Regressed Global Demand Surfaces and investigation of additional predictive environmental and socio-demographic variables.

CHAPTER 7: Conclusion, Overall Contributions, Limitations, and Future Works

7.1 Significant Findings

This thesis examines two PSS Design methodology questions in an effort to resolve difficulties creating RP-based PSS demand estimates. First, *what is the effectiveness of spatially-derived revealed preference data in estimating distributed PSS demand?* This work proposed that estimating binomial distribution parameters *n* (user population size) and *p* (user population product affinity) could predict demand in new situations for distributed PSS. Plots of binomial parameters revealed a continuous surface over the PSS area that allow more accurate prediction of relative ridership levels at PSS locations. This approach was validated with Case Study One, focusing on the 2015 Chicago BSS expansion.

Case Study One provided three pieces of evidence for the efficiency of using binomial parameter estimation for demand prediction. First the algorithm ordering outperformed the operator ordering (rho=.60, stations=46, p<.01), showing a very strong correlation with the ideal ordering (rho=.83, stations=46, p<.01). Secondly, the predicted hourly utilization can also provide insights into the expected average hourly utilization. Algorithm results provide better insight into expected utilization of the new stations (R^2 of 0.43 over 0.38). Finally, when comparing the ideal placement to the recommended algorithm and actual station placement the algorithm placed 94 docks (10% of those added), equivalent to approximately 5,640 pounds of steel, in more desirable locations than the operators did.

The contributions of Case Study One are:

1. The development and application of a revealed preference demand estimation method for distributed product service systems.

2. The creation and application of novel geo-spatial demand surfaces from user data.
3. A tool is created and tested to improve demand prediction for PSS undergoing an increase in system density.

Case Study Two examined the second PSS Design methodology challenge. Case Study Two examined *how designers can compensate for situations where the PSS design environment has changed and limited user data is available to create demand estimations*. The results show that publicly available socio-demographic and environmental variables can be used to create multivariable regressions that estimate the *n* and *p* Demand Surfaces outside of the boundaries previously constrained by available user data.

Case Study Two showed the viability of using multivariable regressions to directly calculate the *n* and *p* surfaces from environmental and socio-demographic variables. The resulting multi-approach algorithm ordering showed a moderate correlation with the ideal ordering (rho=.566, Stations =174, p< .01), while the implemented operator ordering was only weakly correlated (rho=.334, Stations =174, p< .01). When assessing the correlation between the predicted hourly station utilization and observed hourly utilization, the algorithm results in a significantly stronger correlation (R^2=.6, stations=174, p<.01) than the implemented station capacities (R^2=.399, stations=174, p<.01).

Case Study Two made the following contributions:

1. We provide a framework to transform geographically limited available PSS user data into design insights for the portion of the system without user data.
2. A second validation is conducted of the value of *n* and *p* estimations for PSS planning as proposed in the first question examined.
3. We provided a tool for PSS operators planning a system expansion.

4. We identified environmental and socio-demographic variables that correlate with higher Bike Share System usage.

Together, the answer to these two questions provide an initial framework to estimate Reveled Preference demand for many types of PSSs. Next, consolidated recommendations for PSS designers are presented followed by a brief discussion of future work.

7.2 Consolidated Recommendations for PSS Designers

Case Study One resulted in several recommendations for PSS Designers. First, although the Algorithmic approach outperformed traditional methods for the majority of stations, traditional approaches are more successful at identifying the highest performing stations. Thus, distributed PSS operators should identify the highest priority areas utilizing a low-cost, non-deterministic method such as local knowledge, followed by algorithmic assessment. Additionally, algorithmic results can be used as a screening tool for if a station should be added to a location at all. For example, the algorithm identified three stations with a recommended capacity of only one dock (stations 407, 410, 366). The ideal station size for these stations were 1, 0, and 1 dock, thus these dock locations should have been eliminated.

Case Study Two demonstrated the viability of using socio-demographic (income, race, gender) and physical environment (infrastructure, climate, terrain) to directly predict n and p values within the PSS service area. These data sources should be selected after review of current demand estimation approaches and consultation of system experts. Additionally, due to the minimal gains seen when directly implementing user data vs regressions, the minimum data necessary required should be collected prior to shifting to Regressed Global Demand Surfaces. Additional work is required to determine what is the "minimum data necessary."

7.3 Future Works

Although providing a starting framework for PSS demand estimation via n and p estimations, numerous areas for future investigations were identified in this work. Future work should examine the effect of altering the method of n and p estimation upon this approach's effectiveness. Additionally, Case Study One was accurate with a one-year look ahead prediction time, but future investigations should evaluate the time scale for which this approach is effective. Additionally, this approach to RP demand estimation of distributed PSSs should be tested in additional industries and design cases.

In the Case Study Two, the current results can only be applied to a BSS expansion. What about scenarios where only user data from a completely different environment is available? Can the results of this study be used to help BSS operators plan a system installation in a city that currently has no BSS? Does this approach provide a tool for design across markets? These questions will be examined in future work. Additional future work could involve refining a method to combine operator ranking with the results of the Regressed Global Demand Surfaces and investigation of additional predictive environmental and population variables.

Appendix A: Pearson Correlation and Significance Level value for Evaluated Independent Variables

The follow two tables record the level of correlation and associated significance level between the independent variables evaluated in Case Study Two. These were used for multi-collinearity checks when creating the Regressions that formed the basis of the Regressed Global Demand Surfaces. The matrix shows the correlation between the row and column, where the diagonal is all 1.000 due to perfect correlation with itself.

Pearson Correlation

	μ	n	p	AVGn	AVGp	BikePath	Food	Retail	BikeCrash	Park	Train	DISTcbd	DISTlm	St<4800	NearSt	Pop	$	%C	6-64	lowV	Bach	AltCom
μ	1.000	0.147	0.771	0.248	0.785	0.227	0.474	0.339	0.564	-0.538	0.274	-0.552	-0.479	0.543	-0.388	0.176	0.398	0.266	0.239	0.352	0.389	0.361
n	0.147	1.000	-0.241	0.372	0.033	0.130	0.145	0.178	0.064	-0.113	0.087	-0.095	-0.023	0.143	-0.165	-0.007	0.160	0.125	-0.047	0.000	0.063	0.059
p	0.771	-0.241	1.000	0.116	0.728	0.274	0.439	0.280	0.602	-0.422	0.254	-0.541	-0.402	0.556	-0.337	0.150	0.338	0.209	0.197	0.322	0.347	0.328
AVGn	0.248	0.372	0.116	1.000	0.017	0.341	0.188	0.122	0.140	-0.183	0.083	-0.232	-0.052	0.360	-0.292	-0.005	0.384	0.269	-0.120	-0.025	0.106	0.105
AVGp	0.785	0.033	0.728	0.017	1.000	0.277	0.534	0.432	0.727	-0.678	0.298	-0.682	-0.582	0.596	-0.435	0.284	0.418	0.309	0.442	0.504	0.528	0.461
BikePath	0.227	0.130	0.274	0.341	0.277	1.000	0.339	0.265	0.673	-0.301	0.257	-0.766	0.107	0.676	-0.454	-0.080	0.490	0.079	-0.172	-0.059	0.044	0.043
Food	0.474	0.145	0.439	0.188	0.534	0.339	1.000	0.840	0.603	-0.373	0.661	-0.499	-0.296	0.707	-0.496	0.305	0.319	0.229	0.226	0.437	0.452	0.476
Retail	0.339	0.178	0.280	0.122	0.432	0.265	0.840	1.000	0.460	-0.310	0.446	-0.359	-0.229	0.560	-0.456	0.518	0.365	0.423	0.382	0.504	0.621	0.613
BikeCrash	0.564	0.064	0.602	0.140	0.727	0.673	0.603	0.460	1.000	-0.481	0.394	-0.827	-0.211	0.783	-0.502	0.117	0.463	0.196	0.175	0.306	0.314	0.278
Park	-0.538	-0.113	-0.422	-0.183	-0.678	-0.301	-0.373	-0.310	-0.481	1.000	-0.213	0.627	0.257	-0.426	0.362	-0.161	-0.291	-0.267	-0.268	-0.317	-0.311	-0.268
Train	0.274	0.087	0.254	0.083	0.298	0.257	0.661	0.446	0.394	-0.213	1.000	-0.299	-0.158	0.496	-0.305	-0.057	0.169	0.039	-0.089	0.076	0.001	0.079
DISTcbd	-0.552	-0.095	-0.541	-0.232	-0.682	-0.766	-0.499	-0.359	-0.827	0.627	-0.299	1.000	0.124	-0.729	0.531	-0.056	-0.434	-0.123	-0.110	-0.224	-0.251	-0.217
DISTlm	-0.479	-0.023	-0.402	-0.052	-0.582	0.107	-0.296	-0.229	-0.211	0.257	-0.158	0.124	1.000	-0.345	0.297	-0.270	-0.217	-0.096	-0.418	-0.491	-0.463	-0.420
St<4800	0.543	0.143	0.556	0.360	0.596	0.676	0.707	0.560	0.783	-0.426	0.496	-0.729	-0.345	1.000	-0.649	0.224	0.566	0.311	0.094	0.327	0.432	0.441
NearSt	-0.388	-0.165	-0.337	-0.292	-0.435	-0.454	-0.496	-0.456	-0.502	0.362	-0.305	0.531	0.297	-0.649	1.000	-0.236	-0.428	-0.337	-0.179	-0.296	-0.378	-0.378
Pop	0.176	-0.007	0.150	-0.005	0.284	-0.080	0.305	0.518	0.117	-0.161	-0.057	-0.056	-0.270	0.224	-0.236	1.000	0.079	0.414	0.807	0.823	0.896	0.918
$	0.398	0.160	0.338	0.384	0.418	0.490	0.319	0.365	0.463	-0.291	0.169	-0.434	-0.217	0.566	-0.428	0.079	1.000	0.682	-0.031	-0.074	0.301	0.210
%C	0.266	0.125	0.209	0.269	0.309	0.079	0.229	0.423	0.196	-0.267	0.039	-0.123	-0.096	0.311	-0.337	0.414	0.682	1.000	0.205	0.169	0.517	0.469
6-64	0.239	-0.047	0.197	-0.120	0.442	-0.172	0.226	0.382	0.175	-0.268	-0.089	-0.110	-0.418	0.094	-0.179	0.807	-0.031	0.205	1.000	0.854	0.774	0.715
lowV	0.352	0.000	0.322	-0.025	0.504	-0.059	0.437	0.504	0.306	-0.317	0.076	-0.224	-0.491	0.327	-0.296	0.823	-0.074	0.169	0.854	1.000	0.847	0.867
Bach	0.389	0.063	0.347	0.106	0.528	0.044	0.452	0.621	0.314	-0.311	0.001	-0.251	-0.463	0.432	-0.378	0.896	0.301	0.517	0.774	0.847	1.000	0.954
AltCom	0.361	0.059	0.328	0.105	0.461	0.043	0.476	0.613	0.278	-0.268	0.079	-0.217	-0.420	0.441	-0.378	0.918	0.210	0.469	0.715	0.867	0.954	1.000

Significance Level

1.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6-64	0.00	0.00	0.00
0.01	1.00	0.00	0.00	0.57	0.02	0.01	0.00	0.27	0.05	0.13	0.10	0.70	0.01	0.00	0.90	0.01	0.03	0.42	0.99	0.28	0.31
0.00	0.00	1.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.04	1.00	0.77	0.00	0.00	0.03	0.02	0.00	0.15	0.00	0.37	0.00	0.00	0.93	0.00	0.00	0.04	0.66	0.07	0.07
0.00	0.57	0.00	0.77	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.02	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.17	0.00	0.17	0.00	0.31	0.44	0.45
0.00	0.01	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.03	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.27	0.00	0.02	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.13	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.01	0.00	0.00	0.32	0.00	0.50	0.12	0.19	0.99	0.17
0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.00	0.00	0.33	0.00	0.03	0.06	0.00	0.00	0.00
0.00	0.70	0.00	0.37	0.00	0.07	0.00	0.00	0.00	0.00	0.01	0.03	1.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.00
0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.90	0.01	0.93	0.00	0.17	0.00	0.00	0.04	0.01	0.32	0.33	0.00	0.00	0.00	1.00	0.17	0.00	0.00	0.00	0.00	0.00
0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	1.00	0.00	0.60	0.20	0.00	0.00
0.00	0.03	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.50	0.03	0.10	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
0.00	0.42	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.06	0.00	0.10	0.00	0.00	0.60	0.00	1.00	0.00	0.00	0.00
0.00	0.99	0.00	0.66	0.00	0.31	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	1.00	0.00	0.00
0.00	0.28	0.00	0.07	0.00	0.44	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
0.00	0.31	0.00	0.07	0.00	0.45	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

WORKS CITED

[1] Mont, O. K., 2002, "Clarifying the Concept of Product-Service System," J. Clean. Prod., **10**, pp. 237–245.

[2] Baines, T. S., Lightfoot, H. W., Evans, S., Neely, A., Greenough, R., Peppard, J., Roy, R., Shehab, E., Braganza, A., Tiwari, A., Alcock, J. R., Angus, J. P., Bastl, M., Cousens, A., Irving, P., Johnson, M., Kingston, J., Lockett, H., Martinez, V., Michele, P., Tranfield, D., Walton, I. M., and Wilson, H., "State-of-the-Art in Product-Service Systems."

[3] Hankey, S., Lindsey, G., Wang, X., Borah, J., Hoff, K., Utecht, B., and Xu, Z., 2012, "Estimating Use of Non-Motorized Infrastructure: Models of Bicycle and Pedestrian Traffic in Minneapolis, MN," Landsc. Urban Plan., **107**(3), pp. 307–316.

[4] Veryzer, R. W., and De Mozota, B. B., 2005, "The Impact of User-Oriented Design on New Product Development: An Examination of Fundamental Relationships," J. Prod. Innov. Manag., **22**(2), pp. 128–143.

[5] McLoone, H. E., Jacobson, M., Hegg, C., and Johnson, P. W., 2010, "User-Centered Design," Work, **37**(4), pp. 445–456.

[6] He, L., Chen, W., Hoyle, C., Yannou, B., and Paris, E. C., 2012, "Choice Modeling for Usage Context-Based Design," J. Mech. Des., **134**(1343).

[7] Thomas, R. J., 1985, "Problems in Demand Estimation For a New Technology," J. Prod. Innov. Manag., **2**(3), pp. 145–157.

[8] Kang, N., Feinberg, F. M., and Papalambros, P. Y., 2016, "Autonomous Electric Vehicle Sharing System Design," J. Mech. Des., **139**(1), pp. 011402-011402-10.

[9] Fowkes, T., and Preston, J., 1991, "Novel Approaches to Forecasting the Demand for New Local Rail Services," Transp. Res. Part A Gen., **25**(4), pp. 209–218.

[10] ITDP, 2013, "The Bike-Sharing Planning Guide," Inst. Transp. Dev. Policy, p. 152 [Online]. Available: https://www.itdp.org/wp-content/uploads/2014/07/ITDP_Bike_Share_Planning_Guide.pdf. [Accessed: 19-Feb-2018].

[11] Kumar, V. P., and Bierlaire, M., 2012, "Optimizing Locations for a Vehicle Sharing System," *Swiss Transport Research Conference (STRC) (Ascona, Switzerland)*.

[12] Chaudhari, A. M., Zhenghui, S., and Panchal, J. H., 2018, "Analyzing Participant Behaviors in Design Crowdsourcing Contests Using Causal Inference on Field Data," J. Mech. Des., **140**(September).

[13] Wassenaar, H. J., Chen, W., Cheng, J., and Sudjianto, A., 2005, "Enhancing

Discrete Choice Demand Modeling for Decision-Based Design," J. Mech. Des., **127**(4), pp. 514–523.

[14] Ross Morrow, W., Long, M., and MacDonald, E. F., 2014, "Market-System Design Optimization With Consider-Then-Choose Models," J. Mech. Des., **136**(3), pp. 031003-031003-13.

[15] Frischknecht, B. D., Whitefoot, K., and Papalambros, P. Y., 2010, "On the Suitability of Econometric Demand Models in Design for Market Systems," J. Mech. Des., **132**(12), p. 121007.

[16] Williams, N., Azarm, S., and Kannan, P. K., 2008, "Engineering Product Design Optimization for Retail Channel Acceptance," J. Mech. Des., **130**(6), pp. 061402-061402-10.

[17] Wang, Z., Azarm, S., and Kannan, P. K., 2011, "Strategic Design Decisions for Uncertain Market Systems Using an Agent Based Approach," J. Mech. Des., **133**(4), pp. 041003-041003-11.

[18] Chen, H. Q., Honda, T., and Yang, M. C., 2013, "Approaches for Identifying Consumer Preferences for the Design of Technology Products: A Case Study of Residential Solar Panels," J. Mech. Des., **135**(6), p. 061997.

[19] Wang, M., and Chen, W., 2015, "A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering

Design," J. Mech. Des., **137**(7), p. 071409.

[20] Grace Haaf, C., Michalek, J. J., Ross Morrow, W., and Liu, Y., 2014, "Sensitivity of Vehicle Market Share Predictions to Discrete Choice Model Specification," J. Mech. Des., **136**(12), p. 121402.

[21] Kang, C., 2016, "A Simulation Method to Estimate Nonparametric Distribution of Heterogeneous Consumer Preference From Market-Level Choice Data," J. Mech. Des., **138**(12), p. 121402.

[22] Sha, Z., and Panchal, J. H., 2014, "Estimating Local Decision-Making Behavior in Complex Evolutionary Systems," J. Mech. Des., **136**(6), p. 061003.

[23] Kang, N., Feinberg, F. M., and Papalambros, P. Y., 2015, "Integrated Decision Making in Electric Vehicle and Charging Station Location Network Design," J. Mech. Des., **137**(6), p. 061402.

[24] Erkoyuncu, J. A., Roy, R., Shehab, E., and Cheruvu, K., 2011, "Understanding Service Uncertainties in Industrial Product-Service System Cost Estimation," Int. J. Adv. Manuf. Technol., **52**(9–12), pp. 1223–1238.

[25] Dias, G. M., Bellalta, B., and Oechsner, S., 2015, "Predicting Occupancy Trends in Barcelona's Bicycle Service Stations Using Open Data," *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, pp. 439–

445.

[26] Mahony, E. O., and Shmoys, D. B., 2015, “Data Analysis and Optimization for (Citi) Bike Sharing,” Proc. Twenty-Ninth AAAI Conf. Artif. Intell. Data, pp. 687–694.

[27] Froehlich, J., Neumann, J., and Oliver, N., 2009, “Sensing and Predicting the Pulse of the City through Shared Bicycling,” *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1420–1426.

[28] Montoliu, R., 2012, “Discovering Mobility Patterns on Bicycle-Based Public Transportation System by Using Probabilistic Topic Models,” *Advances in Intelligent and Soft Computing*, pp. 145–153.

[29] Simon, H., 1996, *The Sciences of the Artificial, (Third Edition)*, MIT Press, Cambridge.

[30] Davis, M. C., Challenger, R., Jayewardene, D. N. W., and Clegg, C. W., 2014, “Advancing Socio-Technical Systems Thinking: A Call for Bravery,” Appl. Ergon., **45**(2 Part A), pp. 171–180.

[31] Trist, E., 1980, “The Evolution of Socio-Technical Systems: A Conceptual Framework and Action Research Program,” Conf. Organ. Des. Perform., **2**, pp. 1–67.

[32] Hettinger, L. J., Kirlik, A., Goh, Y. M., and Buckle, P., 2015, "Modelling and Simulation of Complex Sociotechnical Systems: Envisioning and Analysing Work Environments," Ergonomics, **58**(4), pp. 600–614.

[33] McDermott, T., Rouse, W., Goodman, S., and Loper, M., 2013, "Multi-Level Modeling of Complex Socio-Technical Systems," *Procedia Computer Science*, pp. 1132–1141.

[34] Emery, F. E., 1959, "Characteristics of Socio-Technical Systems," London Tavistock Inst. Hum. Relations. …, (1959), pp. 1–31.

[35] Vespignani, A., 2012, "Modelling Dynamical Processes in Complex Socio-Technical Systems," Nat. Phys., **8**(1), pp. 32–39.

[36] Moran, P. A. P., 1951, "A Mathematical Theory of Animal Trapping," Biometrika, **38**(3), pp. 307–311.

[37] Draper, N., and Guttman, I., 1971, "Bayesian Estimation of the Binomial Parameter," Technometrics, **13**(3), pp. 667–673.

[38] Byers, A. L., Allore, H., Gill, T. M., and Peduzzi, P. N., 2003, "Application of Negative Binomial Modeling for Discrete Outcomes: A Case Study in Aging Research," J. Clin. Epidemiol., **56**(6), pp. 559–564.

[39] Olkin, I., Petkau, A. J., and Zidek, J. V, 1981, "A Comparison of n

Estimators for the Binomial Distribution A Comparison of n Estimators for the Binomial Distri Buti On," Source J. Am. Stat. Assoc., **76**(375), pp. 637–642.

[40] Feldman, D., and Fox, M., 1968, "Estimation of the Parameter n in the Binomial Distribution Dorian Feldman; Martin Fox," J. Am. Stat. Assoc., **63**(321), pp. 150–158.

[41] Fishman, E., Washington, S., and Haworth, N., 2013, "Bike Share: A Synthesis of the Literature," Transp. Rev., **33**(2), pp. 148–165.

[42] Shaheen, S., Guzman, S., and Zhang, H., 2010, "Bikesharing in Europe, the Americas, and Asia," Transp. Res. Rec. J. Transp. Res. Board, **2143**, pp. 159–167.

[43] National Association of City Transportation Officials, 2016, *Bike Share Station Siting Guide*.

[44] Fishman, E., Washington, S., Haworth, N., and Watson, A., 2014, "Factors Influencing Bike Share Membership: An Analysis of Melbourne and Brisbane," Transp. Res. Part A Policy Pract., **71**, pp. 17–30.

[45] Bullock, C., Brereton, F., and Bailey, S., 2017, "The Economic Contribution of Public Bike-Share to the Sustainability and Efficient Functioning of Cities," Sustain. Cities Soc., **28**, pp. 76–87.

[46] Bonilla Alicea, Ricardo; Watson, Bryan; Tamayo, Laura; Shen, Ziheng; Telenko, C., 2018, “Life Cycle Assessment Comparison of Docking and Dockless Bicycle Share Technologies,” J. Ind. Ecol.

[47] DasGupta, A., and Rubin, H., 2005, “Estimation of Binomial Parameters When Both n, p Are Unknown,” J. Stat. Plan. Inference, **130**(1–2), pp. 391–404.

[48] Hall, P., 1994, “On the Erratic Behavior of Estimators of N in the Binomial N, P Distribution,” J. Am. Stat. Assoc., **89**(425), pp. 344–352.

[49] Tang, V. K. T., Sindler, R. B., and Shirven, R. M., 1987, *Bayesian Estimation of n in a Binomial Distribution*, Alexandria.

[50] “MATLAB Documentation” [Online]. Available: https://www.mathworks.com/help/curvefit/tpaps.html. [Accessed: 14-Dec-2018].

[51] Zhang, J., Pan, X., Li, M., and Yu, P. S., 2016, “Bicycle-Sharing System Analysis and Trip Prediction,” *Proceedings - IEEE International Conference on Mobile Data Management*, pp. 174–179.

[52] Smith, A., 2015, “Crowdsourcing for Active Transportation,” ITE J., **85**(5), pp. 30–35.

[53] O'Brien, O., Cheshire, J., and Batty, M., 2014, "Mining Bicycle Sharing Data for Generating Insights into Sustainable Transport Systems," J. Transp. Geogr., **34**, pp. 262–273.

[54] Fishman, E., Washington, S., and Haworth, N., 2014, "Bike Share's Impact on Car Use: Evidence from the United States, Great Britain, and Australia," Transp. Res. Part D Transp. Environ., **31**, pp. 13–20.

[55] Rudloff, C., and Lackner, B., 2014, "Modeling Demand for Bikesharing Systems," Transp. Res. Rec. J. Transp. Res. Board, **2430**(1), pp. 1–11.

[56] Ome, C., and Latifa, O., 2014, "Model-Based Count Series Clustering for Bike Sharing System Usage Mining : A Case Study with the Vélib System of Paris," ACM Trans. Intell. Syst. Technol., **5**(3), pp. 1–21.

[57] Rixey, R., 2013, "Station-Level Forecasting of Bikesharing Ridership," Transp. Res. Rec. J. Transp. Res. Board, **2387**, pp. 46–55.

[58] 2014, "DIVVY Facebook Post 5 March, 2014" [Online]. Available: https://www.facebook.com/DivvyBikes/photos/a.454681464601866.1073741828.433468746723138/588518824551462/?type=3. [Accessed: 12-Feb-2018].

[59] Dittmar, H., and Ohland, G., 2004, "The New Transit Town: Best Practices in Transit-Oriented Development," Transportation (Amst)., **71**(i), p. 253.

[60] Taylor, W. L., 1964, “Correcting the Average Rank Correlation Coefficient for Ties in Rankings,” J. Am. Stat. Assoc., **59**(307), pp. 872–876.

[61] Wuerzer, T., Mason, S., and Youngerman, R., 2012, *Boise Bike Share Location Analysis*, Boise.

[62] Krykewycz, G., Puchalsky, C., Rocks, J., Bonnette, B., and Jaskiewicz, F., 2010, “Defining a Primary Market and Estimating Demand for Major Bicycle-Sharing Program in Philadelphia, Pennsylvania,” Transp. Res. Rec. J. Transp. Res. Board, **2143**, pp. 117–124.

[63] Wang, X., Lindsey, G., Schoner, J. E., and Harrison, A., 2016, “Modeling Bike Share Station Activity: Effects of Nearby Businesses and Jobs on Trips to and from Stations,” J. Urban Plan. Dev., **142**(1), p. 04015001.

[64] Mateo-Babiano, I., Bean, R., Corcoran, J., and Pojani, D., 2016, “How Does Our Natural and Built Environment Affect the Use of Bicycle Sharing?,” Transp. Res. Part A Policy Pract., **94**, pp. 295–307.

[65] Watson, B. C., and Telenko, C., 2018, “Binomial Parameter Determination and Mapping for Demand Prediction: A Case Study of Bike Sharing Station Expansion Design,” *Proceedings of the ASME 2018 International Mechanical Engineering Congress and Exposition*, ASME, Pittsburgh, PA, pp. 1–12.

[66] Linton, J. D., 2002, "Forecasting the Market Diffusion of Disruptive and Discontinuous Innovation," IEEE Trans. Eng. Manag., **49**(4), pp. 365–374.

[67] Moreno-Ger, P., Torrente, J., Hsieh, Y. G., and Lester, W. T., 2012, "Usability Testing for Serious Games: Making Informed Design Decisions with User Data," Adv. Human-Computer Interact., **2012**, pp. 1–13.

[68] Fischer, G., 2012, "Context-Aware Systems: The 'right' Information, at the 'Right' Time, in the 'Right' Place, in the 'Right' Way, to the 'Right' Person," *Context-Aware Systems: The 'Right' Information, at the 'Right' Time, in the 'Right' Place, in the 'Right' Way, to the 'Right' Person*, Capri Island, Italy.

[69] Nagy, D., Schuessler, J., and Dubinsky, A., 2016, "Defining and Identifying Disruptive Innovations," Ind. Mark. Manag., **57**, pp. 119–126.

[70] Linton, J. D., 2004, "Determining Demand, Supply, and Pricing for Emerging Markets Based on Disruptive Process Technologies," Technol. Forecast. Soc. Change, **71**(1–2), pp. 105–120.

[71] Linton, J. D., and Walsh, S. T., 2001, "Forecasting Micro Electro Mechanical Systems: A Disruptive Innovation," *PICMET '01. Portland International Conference on Management of Engineering and Technology*, Portland.

[72] Bildosola, I., Río-Bélver, R. M., Garechana, G., and Cilleruelo, E., 2017, "TeknoRoadmap, an Approach for Depicting Emerging Technologies,"

Technol. Forecast. Soc. Change, **117**, pp. 25–37.

[73] Altuntas, S., Dereli, T., and Kusiak, A., 2015, "Forecasting Technology Success Based on Patent Data," Technol. Forecast. Soc. Change, **96**, pp. 202–214.

[74] Jeong, Y., Park, I., and Yoon, B., 2016, "Forecasting Technology Substitution Based on Hazard Function," Technol. Forecast. Soc. Change, **104**, pp. 259–272.

[75] Ruan, Y., Hang, C. C., and Wang, Y. M., 2014, "Government's Role in Disruptive Innovation and Industry Emergence: The Case of the Electric Bike in China," Technovation, **34**(12), pp. 785–796.

[76] "Chicago Data Portal," Chicago Data Portal [Online]. Available: https://data.cityofchicago.org/. [Accessed: 24-Feb-2019].

[77] Belsley, D. A., Kuh, E., and Welsch, R. E., 1980, *Regression Diagnostics*, John Wiley and Sons, Hoboken.